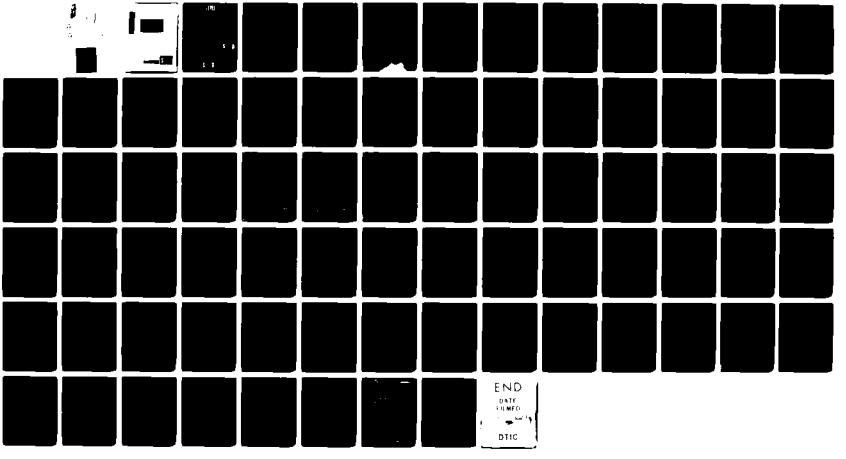
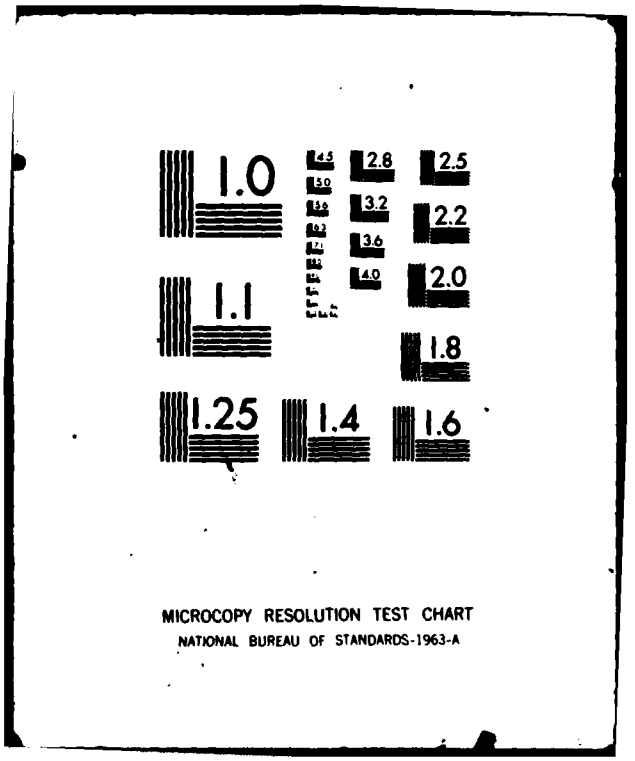
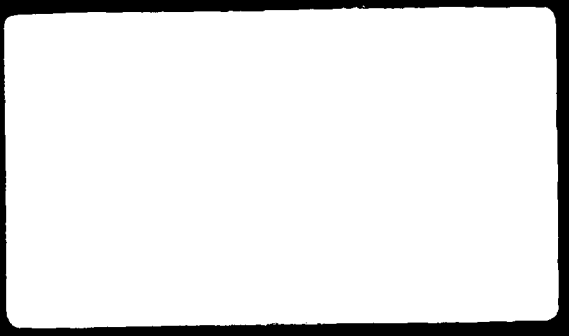
AD-A097 536 CLEMSON UNIV SC DEPT OF MATHEMATICAL SCIENCES F/G 12/1
ON THE DEGREE OF INFLATION OF MEASURES OF FIT INDUCED BY EMPIRI--ETC(U)
AUG 80 T B EDWARDS, K T WALLENIUS N00014-75-C-0451
UNCLASSIFIED N120 NL



END
DATE FILMED
DTIC




MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A







formulas for the parameters based on the beta function as N becomes large. An upward bias results when these formulas are applied to the Monte-Carlo data. The authors attribute

LEVEL II

(6)

# ON THE DEGREE OF INFLATION OF MEASURES OF FIT INDUCED BY EMPIRICAL MODEL BUILDING.

(10) T.B. Edwards and K.T. Wallenius
Clemson University



Department of Mathematical Sciences

(9) Technical Report #350
(11) August 1980
(14) N12Ø, TR-35Ø
(12) 76



This work was supported in part by the Office of Naval Research under Contract N00014-75-C-0451.
(15)



407183 mt

## ABSTRACT

This dissertation explores the distributional properties of commonly used statistics developed in the course of empirical model building. A review of some of the more noteworthy efforts to investigate the distribution of the coefficient of determination, $R^2$, in best subset regression is given. To overcome the shortcomings of these results, a permutation test based on Fisher's randomization test is developed to provide a practical basis for assessing the statistical significance of a regression in such situations.

An investigation is made into the distributional properties of the multiple correlation coefficient in the choice of a transformation of the dependent variable, y. The study investigates the possibility that pedestrian use of transformations, such as $y^* = (y+c)^p$, may lead to an inflationary effect on the sample correlation.

A practical management science application of the statistical procedures developed in this study is explored in the area of parametric cost estimation.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER I

## INTRODUCTION

The most widely used and abused data analytic methodology is regression analysis (4). Many books, notably (7), (9) and (15), and thousands of research papers attest to the popularity and importance of these powerful statistical procedures. The advent of high-speed digital computers and associated statistical software packages has made regression analysis accessible to users in all fields of research. In particular the new technological developments in time-shared computing literally bring these and other procedures into the manager's office providing the means for assessing decision alternatives at a moment's notice. Sophisticated techniques, now routinely applied, were impractical only 20 years ago because of enormous computational requirements.

For some methodologies, in particular empirical model building, statistical theory is not keeping pace with ever-expanding computational capabilities in the sense that data analysts are developing and using algorithms which lead to results whose statistical properties are not fully understood. This statement is not intended as a criticism of exploratory data analysis per se, but it does identify an area of practical significance whose theoretical foundation is shaky at best.

Unlike confirmatory statistical techniques (such as hypothesis testing), wherein inferences are made within the framework of a given model, the term "empirical model building" is used to describe the process of "letting the data speak for itself." In searching for possible relationships among a collection of variables, the data analyst may allow the sample data to answer such questions as, "Which variables should be included in the model?" and "What model structures should be contemplated?"

The purpose of this dissertation is to explore the distributional properties of commonly used statistics developed in the course of empirical model building. Theoretical results are obtained in certain tractable cases. Simulation is employed to develop insight in those situations where explicit mathematical results have been elusive.

It is well known that the use of empirical variable selection techniques in multiple regression leads to inflated values of the coefficient of determination, $R^2$. The degree of this inflation is not well understood. What makes the problem difficult is the fact that the distribution of $R^2$ depends not only on the underlying relationship among the variables, but on the data analytic tools used to develop the model. Attempts have been made to obtain approximations and asymptotic results for special cases of this problem. Chapter II reviews some of the more noteworthy efforts to investigate the distribution of $R^2$ in best

subset regression. Best subset regression is concerned with the problem of determining the subset of size k out of p candidate predictor variables which maximizes some function of $R^2$, where k itself may be data-dependent. Special attention is allotted an asymptotic result of Alam and Wallenius (2). A proof of their result is provided. Since their asymptotic distribution of $R^2$ is derived by allowing the sample size to grow large, an investigation is performed to determine an appropriate sample size for an adequate approximation.

The approximations, alluded to above provide insight but little help of a practical nature in testing for statistical significance of the sample $R^2$ resulting from data analytic selection techniques. Chapter III addresses this problem. The shortcomings of the classical statistical tests for these situations are reviewed. In order to overcome these limitations, a new approach is introduced which yields an exact test conditioned on the sample data and selection technique. This test is most useful in situations where the number of observations is small compared to the number of candidate predictor variables. In particular, this test is valid if the number of potential predictor variables exceeds the available degrees of freedom. The classical F test cannot be used in this case. Determining the power of this test is a difficult problem and remains unsolved. A simulation is used to compare the power of the new test to

that of the classical F test for several cases where the latter is valid.

Chapter IV deals with the distribution of the multiple correlation coefficient in multiple regression when the data is used to determine the choice of a transformation of the dependent variable y. The family of transformations considered is of the form $y^* = (y+c)^p$. This is a widely used family of transformations of practical importance. It is often used, as Tukey (19) puts it, "to remove apparent ills from the data ... aiding in the analysis by bending the data nearer the Procrustean bed of the assumptions underlying conventional analysis." The data is employed to determine c and p in such a way that the relationship between $y^*$ and a single predictor x is more nearly linear than that between y and x. The study investigates the possibility that pedestrian use of this transformation may lead to an inflationary effect on the sample correlation. The results indicate some interesting phenomena which are illustrated in examples and lead to a theorem.

Chapter V explores a practical management science application of the statistical procedures developed in this study. A problem often faced by costing and pricing analysts involves estimating the cost of a proposed system. One approach to this problem is independent parametric cost estimation. A description of this method, its advantages, and disadvantages are given. Actual cost and performance

data obtained from the Navy Weapons Center, China Lake, California, are analyzed using the methodology of Chapter III.

Chapter VI contains a discussion of the inherent difficulties of empirical model building and identifies some areas in which further research is required.

# CHAPTER II

## APPROXIMATIONS OF THE DISTRIBUTION OF $R^2$ IN BEST SUBSET REGRESSION

In recent years a great deal of interest has been expressed in the distributional properties of $R^2$ and other statistical measures of fit for regression models when variable selection techniques are employed. Historically, Fisher (10) derived the general sampling distribution of $R^2$ when sampling from a multivariate normal distribution. The distribution theory for the sample $R^2$ statistic in empirical model building is quite complex. The difficulty stems from the fact that the distribution depends not only on the underlying relationship among the variables but also on the variable selection criterion.

This chapter reviews some interesting approximating formulas and asymptotic results for the distribution of $R^2$ in best subset regression. Here the term "best subset regressions" refers to the following situations. The data analyst has a set of n independent observations on p candidate predictor variables and one dependent variable. The goal of the analysis is to determine the k-variable regression equation which maximizes the sample coefficient of determination for various values of k. The difficulty with this analysis is assessing the statistical significance of $R^2$ for a given value of k.

Diehr and Hoflin (8) utilize a Monte-Carlo approach to devise a function purported to estimate the distribution of the sample $R^2$ in best subset regression for samples selected from a p+1 dimensional multivariate normal population with zero mean and identity covariance matrix. Monte-Carlo estimates of the $(1-\alpha)$ percentile points, $R^2(k,p,n,\alpha)$, of the sample distribution are obtained for selected values of k, p, and n. This is accomplished by generating 100 samples of size n from the null distribution and determining and saving the maximum $R^2$ associated with the best k-variable regression equation for k from 1 to p. The set of 100 $R^2$ values corresponding to a particular collection of k, p, and n values are ordered to give estimates of the percentage points. By visually examining some of the Monte-Carlo results, the authors note that a function of the form

$$\hat{R}^2(k,p,n,\alpha) = w(1-v^k)$$

seems to provide a reasonable fit of the simulation results when w and v are determined from the known boundary values $R^2(1,p,n,\alpha)$ and $R^2(p,p,n,\alpha)$. The authors suggest that a statistical test based on this formula is an improvement over the standard tests for the empirical researcher since the number of independent variables which has been searched is taken into account. This test, while more appropriate than the F test in spirit at least, would serve only to give insight into the results. The nominal "significance level"

is somewhat suspect since the percentile points are based on an ad hoc fit of a Monte-Carlo distribution.

Rencher and Pun Fu-Ceayong (16) extend the results of Diehr and Hoflin by computing the mean of the inflated $R^2$ under best subset selection, allowing for correlated predictor variables, and including the situation where the number of candidate predictor variables exceeds the number of observations.

As expected, their Monte-Carlo study indicates that the inflation of $R^2$ is somewhat less when the predictor variables are intercorrelated. To supplement Monte-Carlo estimates for the mean and percentage points of the distribution of $R^2$ under selection, the authors obtain asymptotic approximations for these parameters. For a k-variable model without selection, $R^2$ has a beta distribution in the null case (10). Thus, the distribution function of $R^2$ is given by

$$F_{a,b}(R^2) = \beta(R^2;\ a,b)/\beta(1;\ a,b)$$

where, for $0 \leq x \leq 1$,

$$\beta(x;\ a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt,$$

$a = k/2$ and $b = (n-k-1)/2$.

The number of possible k-variable prediction equations is $N = p!/[k!(p-k)!]$. By assuming the corresponding N values of $R^2$ are independent, the authors obtain asymptotic

formulas for the parameters based on the beta function as N becomes large. An upward bias results when these formulas are applied to the Monte-Carlo data. The authors attribute this bias to the assumption of independence noted above. The formulas are modified to correct for this biasedness by adjusting the value of N via a function of the form $(\ell n\ N)^{cN^d}$, where c and d are empirically determined from the Monte-Carlo results. Their final approximating formulas for the mean and $\gamma$-th percentile of the distribution $R^2$ are, respectively,

$$\hat{E}(R^2) = 1-F^{-1}_{b,a}[1/(\ell n\ N)^{1.5N^{.04}}]\,\Gamma(1+1/w)$$

and

$$\hat{R}^2_\gamma = F^{-1}_{a,b}[1 + \ell n\ \gamma/(\ell n\ N)^{1.8N^{.04}}]$$

where $w = (n-k-1)/2$.

It is suggested that these formulas can be used as possible guidelines for assessing the significance of $R^2$ values obtained in best subset regression applications. However, the empirical researcher might feel that his confidence in their use is overshadowed by their ominous appearance and computational complexity.

Zirphile (23) derives an asymptotic approximation for the $(1-\alpha)$ percentiles of the distribution of $R^2$ in best subset regression, as the sample size n is made large, using extreme value theory. This approximating formula gives

percentage points which are as large as 1.5 for some small values of n (16). This poor performance may be due in part to the fact that his results are based on the assumption that under the null hypothesis the asymptotic distribution of $R^2$ for a k-variable equation without selection is normal. The actual distribution does not tend to normality for large n under the null hypothesis (10) so that Zirphile's results are of dubious value.

Alam and Wallenius (2) derive a very interesting asymptotic result in the following

Theorem: Let $\underline{Z}' = (Y, X_1, X_2, \ldots, X_p)$ have a (p+1)-variate normal distribution with arbitrary mean vector $\underline{\mu}$ and diagonal covariance matrix $\underline{\underline{\Sigma}}$. Given a sample of size n on $\underline{Z}$, let $r_{i_1, i_2, \ldots, i_k}$ denote the sample multiple correlation coefficent between y and the k predictor variables $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ where $1 \leq k \leq p$. Let $r_n^2(p,k)$ denote the maximum of all $\binom{p}{k}$ values of $r_{i_1, i_2, \ldots, i_k}$. Then as n tends to infinity, $(n-1)\, r_n^2(p,k)$ converges (with probability 1) to a random variable $r^2(p,k)$ distributed as the sum of the k largest order statistics of a random sample of size p from a chi-square distribution with one degree of freedom.

Proof: Let $\underline{X} = (X_1, X_2, \ldots, X_{p+1})' \sim MVN(\underline{\mu}, \underline{\underline{\Sigma}})$ and assume $\underline{\underline{\Sigma}}$ is of the form

$$\underline{\Sigma} = \begin{bmatrix} \sigma_{11} & 0 & 0 & \cdots & 0 \\ 0 & \sigma_{22} & & & \\ 0 & & & & \\ \vdots & & & & \\ 0 & & & & \sigma_{p+1,p+1} \end{bmatrix}$$

We may assume, without loss of generality, that $\underline{\mu} = 0$. Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ be a random sample size n on $\underline{X}$. Let

$$\underline{A} = \sum_{i=1}^{n} \sum_{j=1}^{n} (\underline{x}_i - \bar{\underline{x}})(\underline{x}_j - \bar{\underline{x}})'$$

where

$$\bar{\underline{x}} = 1/n \sum_{j=1}^{n} \underline{x}_j .$$

Next, partition $\underline{A}$ as

$$\underline{A} = \begin{bmatrix} a_{11} & \underline{A}_{12} \\ \underline{A}_{21} & \underline{A}_{22} \end{bmatrix}$$

and let $\underline{y}' = (x_{11}, x_{12}, \ldots, x_{1n})$ be the first component of each sample observation $\underline{x}_i$, $i = 1,2,\ldots,n$. Consider the conditional distribution of

$$\underline{A}_{21} = \begin{pmatrix} a_{21} \\ a_{31} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{(p+1)1} \end{pmatrix}$$

given $\underline{Y} = \underline{y}$.

The matrix

$$\underline{\underline{X}} = \begin{pmatrix} y_1 & y_2 & \cdots & y_j & \cdots & y_n \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2n} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ x_{(p+1)1} & x_{(p+1)2} & \cdots & x_{(p+1)j} & \cdots & x_{(p+1)n} \end{pmatrix}$$

is column-wise independent and the jth column is conditionally

$$N(\frac{\underline{\Sigma}_{21}}{\sigma_{11}} y_j, \underline{\underline{\Sigma}}_{22} - \underline{\Sigma}_{21} \sigma_{11}^{-1} \underline{\Sigma}_{12}).$$

However, since we are assuming $\underline{\Sigma}_{21} = 0$, we have

$$\left[\begin{pmatrix} x_{2j} \\ x_{3j} \\ \vdots \\ x_{(p+1)j} \end{pmatrix} \mid \underline{Y} = \underline{y}\right] \sim N(\underline{0}, \underset{=}{\Sigma}_{22}).$$

Since $a_{i1} = \sum_{j=1}^{n} (y_j - \bar{y}) x_{ij}$, it follows that

i) $E(a_{i1} | \underline{Y} = \underline{y}) = 0.$

ii) $\operatorname{cov}(a_{i1}, a_{k1} | \underline{y}) = \sum_{\ell=1}^{n} \sum_{j=1}^{n} (y_j - \bar{y})(y_\ell - \bar{y}) \operatorname{cov}(x_{ij}, x_{k\ell} | \underline{y}).$

Note that $\operatorname{cov}(x_{ij}, x_{k\ell} | \underline{y}) = 0$ unless $j = \ell$. Thus

$$\operatorname{cov}(a_{i1}, a_{k1} | \underline{y}) = \sum_{j=1}^{n} (y_j - \bar{y})^2 \operatorname{cov}(x_{ij}, x_{kj} | \underline{y})$$

$$= a_{11} \sigma_{ik}.$$

So, given $\underline{Y} = \underline{y}$, $\underline{A}_{21} \sim N(0, a_{11} \underset{=}{\Sigma}_{22})$. Recall that

$$\hat{\rho}^2 = \frac{\underline{A}_{12} \underset{=}{A}_{22}^{-1} \underline{A}_{21}}{a_{11}} = \frac{\underline{A}_{12} \underset{=}{\Sigma}_{22}^{-1/2}}{\sqrt{a_{11}}} \underset{=}{\Sigma}_{22}^{1/2} \underset{=}{A}_{22}^{-1} \underset{=}{\Sigma}_{22}^{1/2} \frac{\underset{=}{\Sigma}_{22}^{-1/2} \underline{A}_{21}}{\sqrt{a_{11}}}$$

and let

$$\underline{z} = \frac{\underset{=}{\Sigma}_{22}^{-1/2} \underline{A}_{21}}{\sqrt{a_{11}}}.$$

Then, given $\underline{Y} = \underline{y}$,

1. $\underline{z} \sim N(\underline{0}, \underline{I})$

2. $\hat{\rho}^2 = \underline{z}' \underline{\Sigma}_{22}^{1/2} \underline{A}_{22}^{-1} \underline{\Sigma}_{22}^{1/2} \underline{z}$.

Thus, $(n-1)\hat{\rho}^2 = \underline{z}' \underline{\Sigma}_{22}^{1/2} (\frac{1}{n-1} \underline{A}_{22})^{-1} \underline{\Sigma}_{22}^{1/2} \underline{z}$.

Since $(\frac{1}{n-1} \underline{A}_{22})^{-1} \overset{a.s.}{\to} \underline{\Sigma}_{22}^{-1}$, we have for large n, given $\underline{Y} = \underline{y}$,

$$(n-1)\hat{\rho}^2 \overset{a.s.}{\to} \underline{z}'\underline{z} = \sum_{i=2}^{p+1} z_i^2 \sim \chi^2(p) \text{ Vy}.$$

That is, the asymptotic conditional distribution does not depend on the conditioning value of $\underline{Y}$. Therefore, the distribution of $(n-1)\hat{\rho}^2$, for large n, is chi-square with p degrees of freedom.

Suppose we wish to consider all $\binom{p}{k}$ subsets of size k from the set $\{x_2, x_3, \ldots x_{p+1}\}$ of predictor variables and compute the sample multiple correlation coefficient between y and each such subset. Let $R_{i_1,i_2,\ldots,i_k}$ denote the multiple correlation between y and the set $\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$ and let

$$R(k) = \max_{\{i_1,i_2,\ldots,i_k\}} R_{i_1,i_2,\ldots i_k}.$$

From above, we have for large n

$$(n-1)R^2_{2\ldots(p+1)} = z_2^2 + z_3^2 + \ldots + z_{p+1}^2$$

where

$$z_i = \frac{a_{i1}}{\sqrt{a_{11}}\sqrt{\sigma_{ii}}} \sim N(0,1).$$

Thus

$$(n-1)R^2_{2\ldots(p+1)} = \frac{1}{a_{11}}\left(\frac{a_{21}^2}{\sigma_{22}} + \frac{a_{31}^2}{\sigma_{33}} + \ldots + \frac{a_{(p+1)1}^2}{\sigma_{p+1,p+1}}\right)$$

for large n. More generally, we see that

$$(n-1)R^2_{i_1 i_2 \ldots i_k} = \frac{1}{a_{11}}\left(\frac{a_{i_1 1}^2}{\sigma_{i_1 i_1}} + \frac{a_{i_2 1}^2}{\sigma_{i_2 i_2}} + \ldots + \frac{a_{i_k 1}^2}{\sigma_{i_k i_k}}\right)$$

so that

$$(n-1)R^2(k) = \max_{\{i_1, i_2, \ldots, i_k\}} (n-1)R^2_{i_1 i_2 \ldots i_k}$$

$$= \max_{\{i_1, i_2, \ldots, i_k\}} \left(\frac{a_{i_1 1}^2}{a_{11}\sigma_{i_1 i_1}} + \frac{a_{i_2 1}^2}{a_{11}\sigma_{i_2 i_2}} + \right.$$

$$\left. \ldots + \frac{a_{i_k 1}^2}{a_{11}\sigma_{i_k i_k}}\right).$$

Note that $\left\{\frac{a_{j1}^2}{a_{11}\sigma_{jj}} : j = 2,3,\ldots,p+1\right\}$ represents a random sample of size p from a $\chi^2(1)$. Thus $(n-1)R^2(k)$ is the sum

of the k largest values of $(\frac{a_{j1}^2}{a_{11}\sigma_{jj}})$ or the sum of the k largest order statistics of a random sample of size p from a chi-square distribution with one degree of freedom. Q.E.D.

This theorem is of particular interest since it provides information about the degree of inflation of the F statistic in best subset regression as a corollary. Under the null hypothesis for a particular k-variable model, the F statistic is distributed as a constant times the quotient of independent $\chi^2$ random variables, that is

$$F \sim \frac{\chi^2(k)/k}{\chi^2(n-1-k)/(n-1-k)} .$$

For large n, by Theorem 20.6 of (6) the denominator converges in probability to 1. Thus, F converges in distribution to a random variable distributed as a $\chi^2(k)/k$ or, equivalently, as the average of k independent observations from a chi-square distribution with one degree of freedom. If the best subset of size k out of p predictors is selected, the associated F statistic is given by

$$F_{max} = \frac{r_n^2(p,k)}{1-r_n^2(p,k)} \left(\frac{n-1-k}{k}\right) = \frac{1}{k}\left[\frac{(n-1)r_n^2(p,k)}{1-r_n^2(p,k)}\right] - \frac{r_n^2(p,k)}{1-r_n^2(p,k)} .$$

Since $r_n^2(p,k)$ converges to zero as a direct result of the Alam-Wallenius theorem, $F_{max}$ converges in distribution

to $\frac{1}{k}r^2(p,k)$ which is distributed as $\frac{1}{k}\sum_{i=1}^{k}\chi^2_{[p-i+1]}(1)$, the average of the k largest of p independent observations from a chi-square distribution with one degree of freedom. This comparison of F and $F_{max}$ gives the clearest picture of the nature of the inflation of $R^2$ in best subset regression.

The sample size necessary for an adequate approximation by an asymptotic methodology is always of prime concern. An investigation into this question is made by means of a Monte-Carlo approach for the Alam-Wallenius result. A simulation is performed in a straight-forward manner to compare the distributions of the two statistics involved. A random sample of size p is selected from a chi-square distribution with one degree of freedom. The k largest values in the sample are added together to yield one observation on the statistic $r^2(p,k)$. This procedure is repeated 1000 times, and a relative frequency histogram is developed. Figure 1 shows such a histogram for $r^2(8,4)$.

Alam and Wallenius (3) show that the distribution function of statistic $r^2(p,k)$ can be expressed as an infinite linear combination of gamma distribution functions. The shape of Figure 1 resembles the gamma density. For these reasons, an attempt is made to fit a gamma density to the simulation results. Recall that the gamma density is a two-parameter function which may be written as

$$f(x) = \begin{cases} \dfrac{1}{\beta^{\alpha}\,\Gamma(\alpha)}\; x^{\alpha-1}\; e^{-x/\beta} & x > 0 \\ \\ 0 & \text{otherwise} \end{cases}$$

The method of moments is applied to the data presented in Figure 1 to obtain estimates of $\alpha$ and $\beta$: $\hat{\alpha} = 3.6$ and $\hat{\beta} = 2.0$. Figure 2 depicts the histogram of $r^2(8,4)$ with this gamma density superimposed. Table I gives percentage points for the sample data and this gamma density.

TABLE I.

Percentage Points of $r^2(8,4)$ and the Gamma Density

| | 90% | 95% | 99% |
|---|---|---|---|
| $r^2(8,4)$ | 12.15 | 14.18 | 18.91 |
| Gamma | 12.29 | 14.36 | 18.80 |

The distribution of $r^2(8,4)$ appears very similar to that of a random variable distributed as a gamma with $\alpha = 3.6$ and $\beta = 2$. It appears that the infinite sum mentioned above may be dominated by a single gamma distribution function.

To obtain an empirical distribution for the statistic $(n-1)r_n^2(p,k)$, a random sample of size n is selected from a p+1 dimensional multivariate normal population with zero

mean and identity covariance matrix. The best k-variable regression equation is then determined by use of an efficient search of the $\binom{p}{k}$ possible regressions, and the resulting value of $(n-1)r_n^2(p,k)$ is saved. This process is repeated 500 times resulting in relative frequency distributions of $(n-1)r_n^2(p,k)$. For $p = 8$ and $k = 4$, Figures 3-6 depict histograms of these frequencies for $n = 10$, 25, 50 and 100, respectively, with the superimposed density of the gamma distribution.

The results of this simulation offer no definitive answer to the question of appropriate sample size. However, it is possible to form some conclusions after visually examining the histograms. In a hypothesis testing framework, the right-tail of the distributions will be important in the decision making process. The statistic $r^2(p,k)$ seems to overestimate the probability in the right-tail for small values of n as can be seen in Table II. A statistical test based on this distribution would appear to be a conservative test for small values of n in that the actual significance level is less than the nominal level.

TABLE II.

Percentage Points of $r^2(8,4)$, Gamma and $(n-1)r_n^2(8,4)$

| | 90% | 95% | 99% |
|---|---|---|---|
| $r^2(8,4)$ | 12.15 | 14.18 | 18.91 |
| gamma | 12.29 | 14.36 | 18.80 |
| 9 $r_{10}^2(8,4)$ | 7.85 | 8.14 | 8.69 |
| 24 $r_{25}^2(8,4)$ | 9.60 | 10.54 | 13.13 |
| 49 $r_{50}^2(8,4)$ | 10.39 | 11.82 | 15.46 |
| 99 $r_{100}^2(8,4)$ | 10.41 | 11.91 | 15.68 |

The approximations presented in this chapter show the approaches that have been used to explore the distributional properties of $R^2$ under best subset regression. These results provide a better understanding but afford little help of a practical nature in testing for statistical significance of the sample $R^2$ resulting from data analytic selection techniques. In the next chapter, a new and exact method to deal with this important problem is developed.

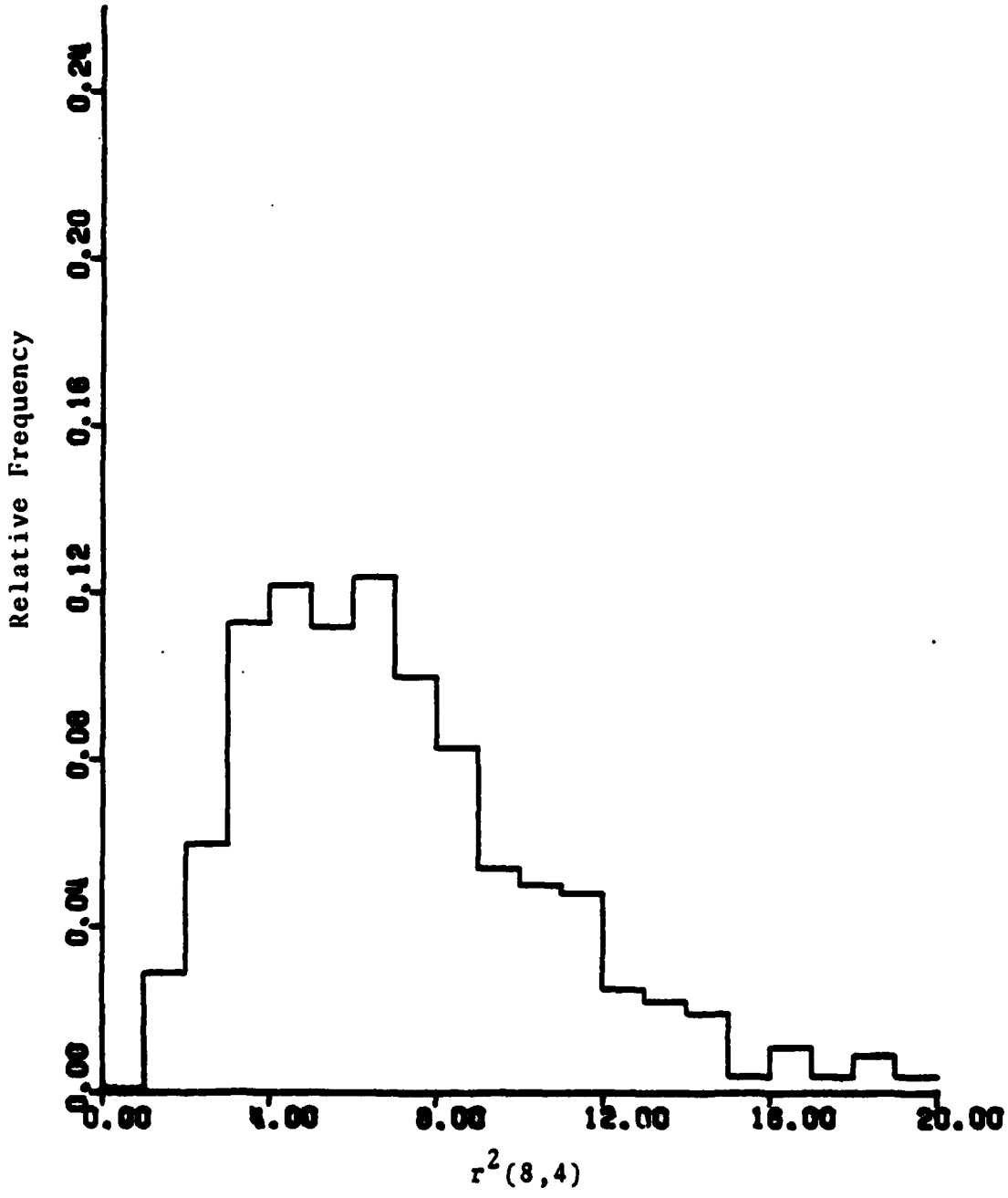



Figure 1: Relative Frequency of $r^2(8,4)$

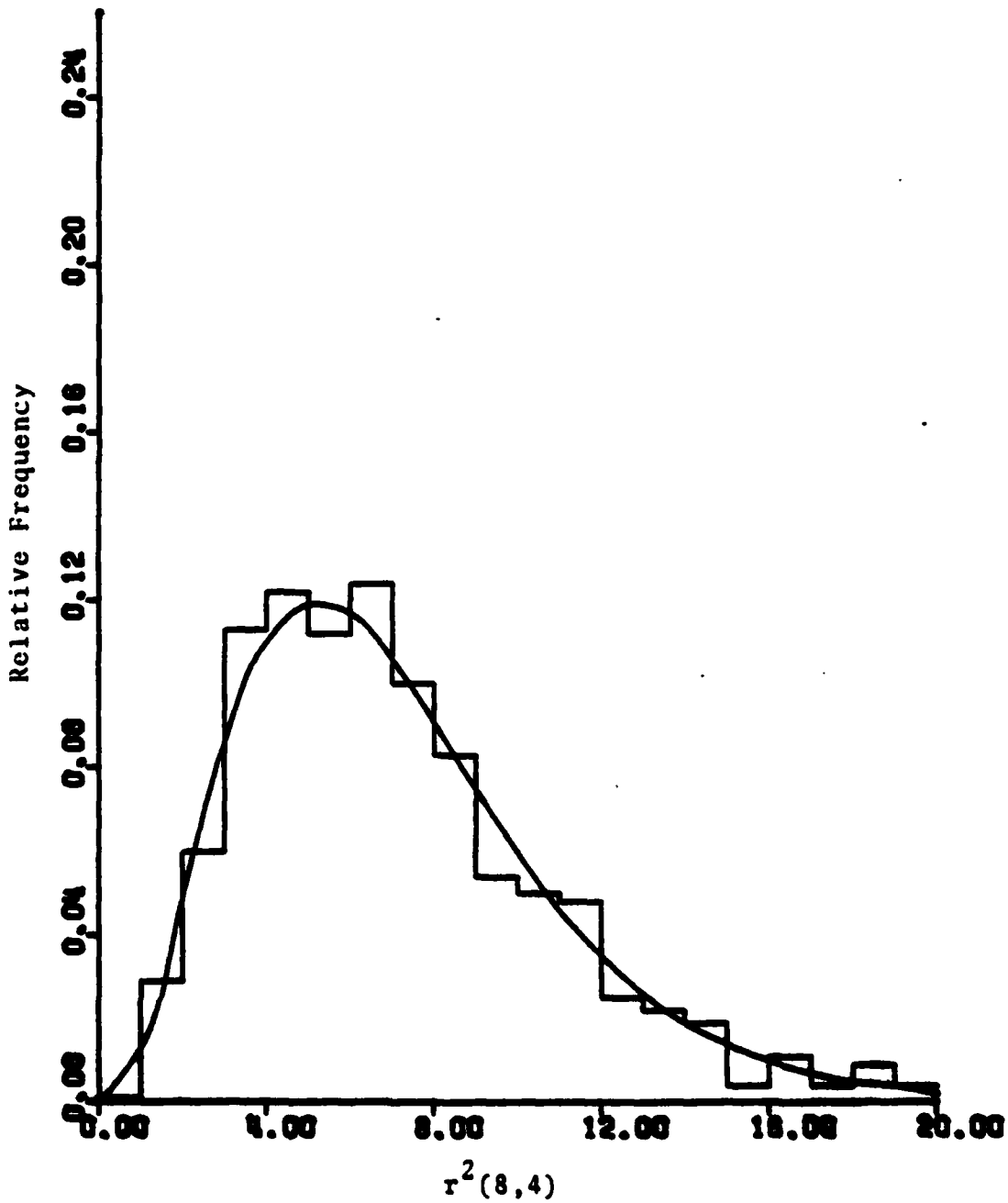



Figure 2: Relative Histogram of $r^2(8,4)$ and Gamma Density

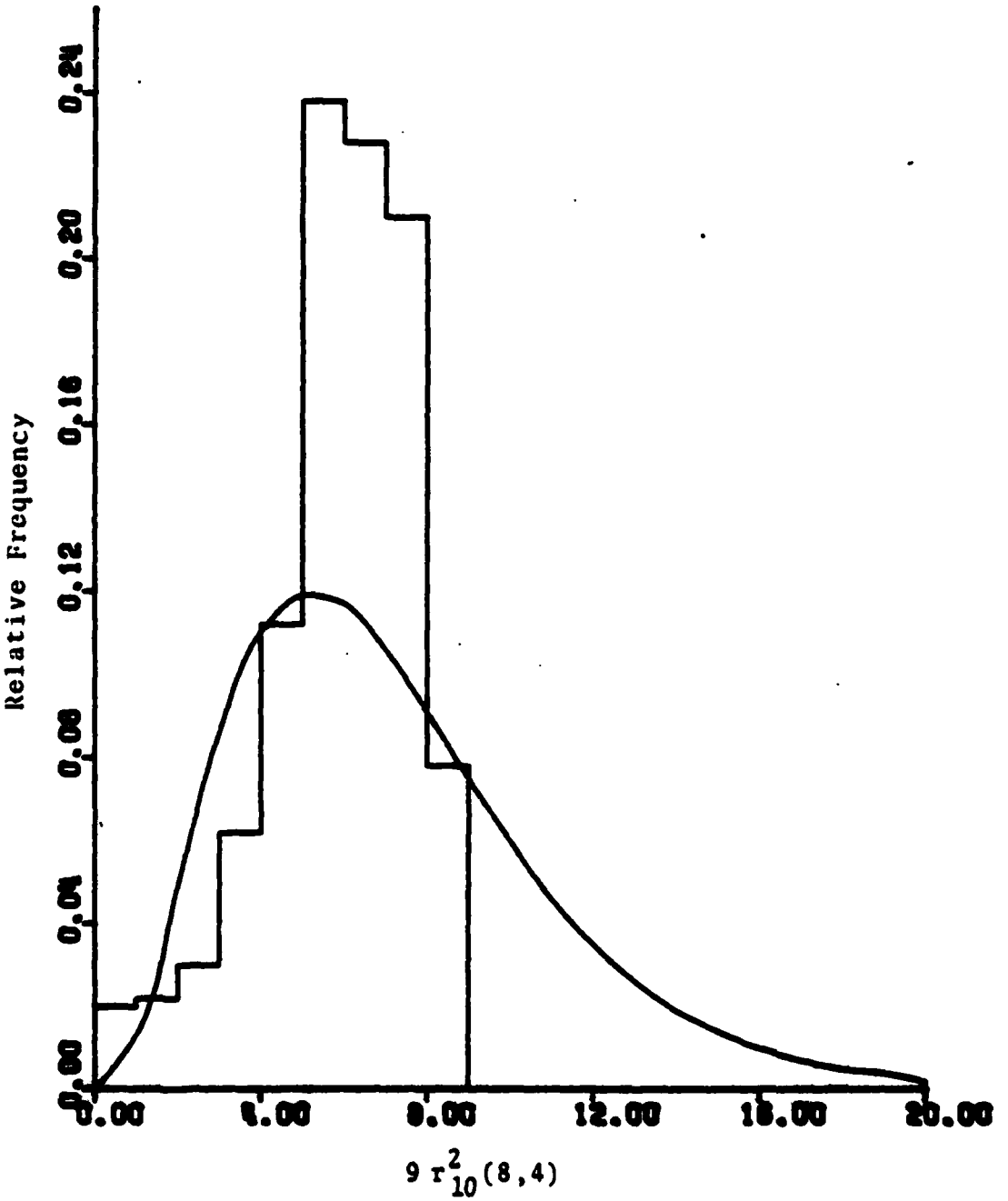



Figure 3: Relative Frequency of $9\, r_{10}^2(8,4)$

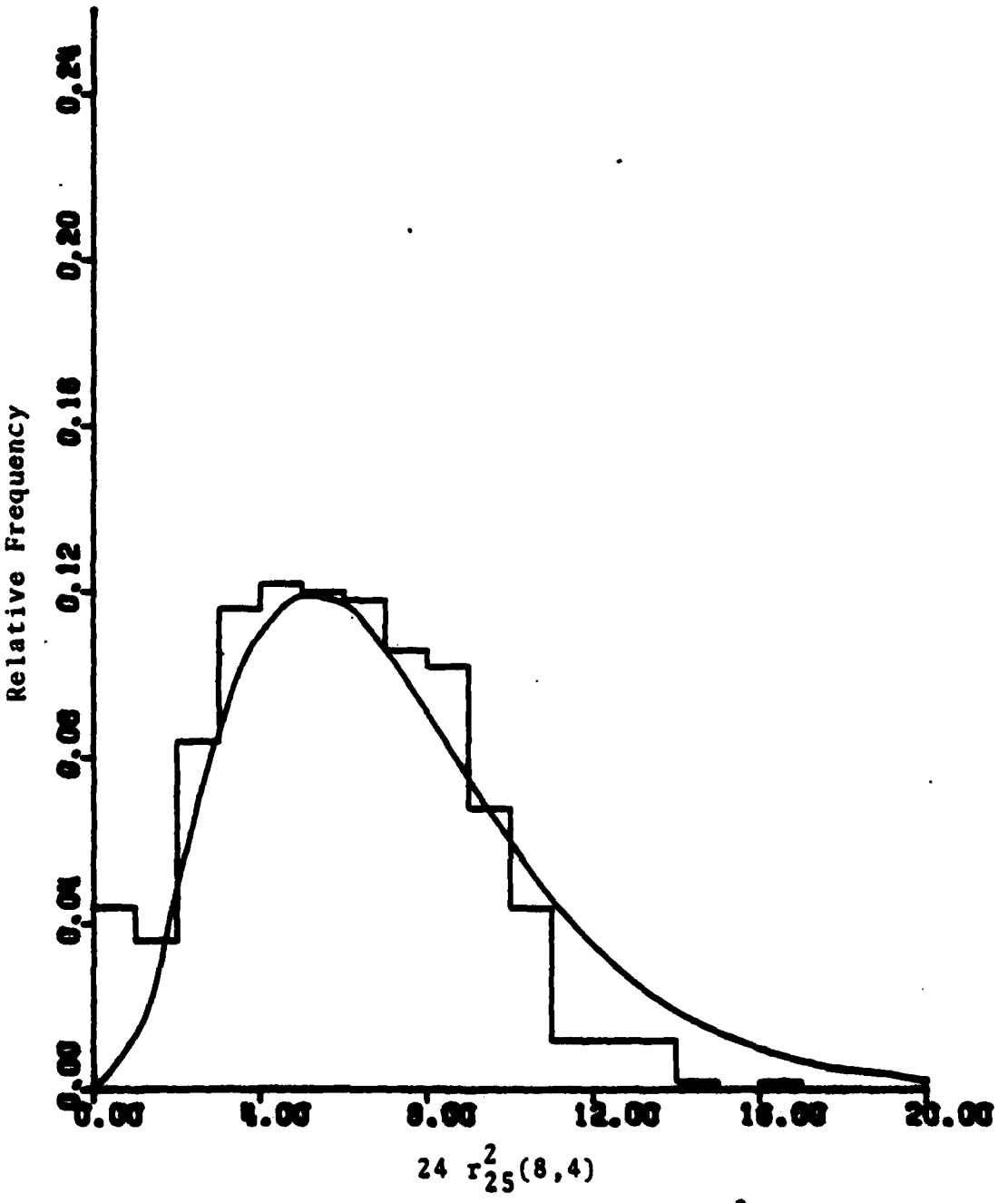



Figure 4: Relative Freqency of 24 $r_{25}^2(8,4)$

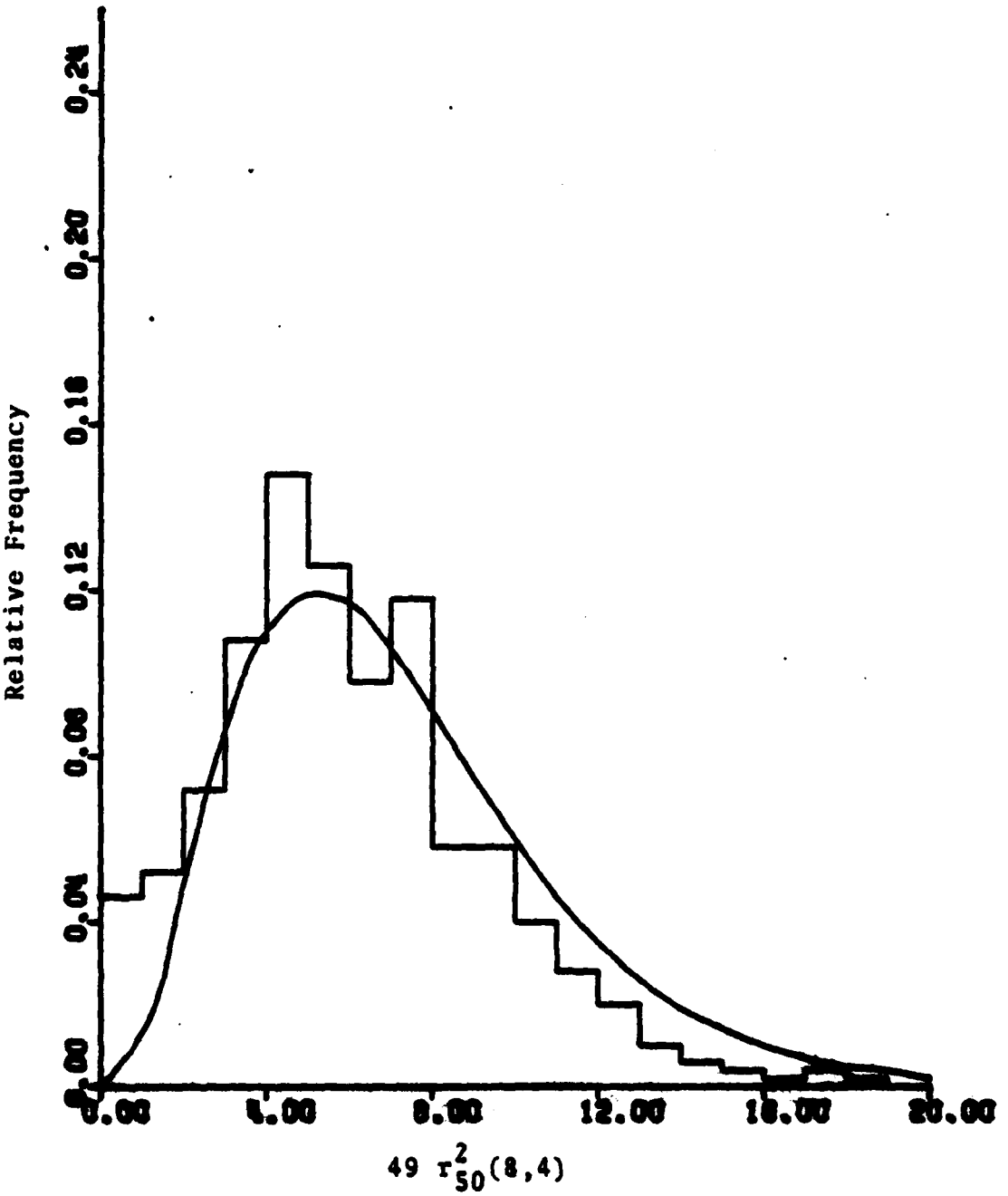



Figure 5: Relative Frequency of 49 $r_{50}^2(8,4)$

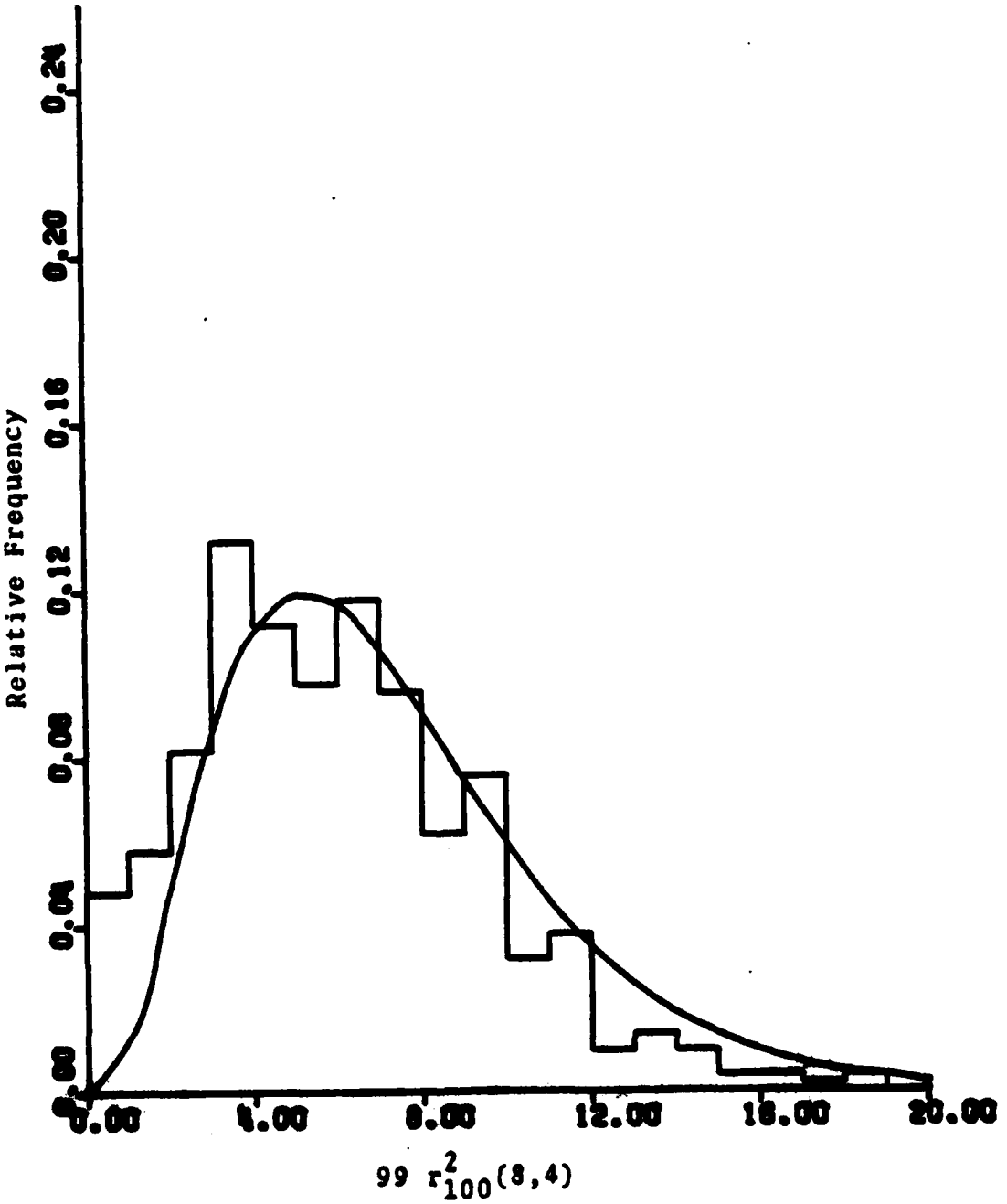



Figure 6: Relative Frequency of 99 $r^2_{100}(8,4)$

# CHAPTER III

## SIGNIFICANCE TESTS AND TESTS OF MODELS IN SUBSET REGRESSION

In Chapter II some asymptotic approximations for the null distribution of $R^2$ in the case where the predictor variables are orthogonal were discussed. In this chapter, the focus is on exact statistical tests for the finite sample size case and the result is generalized to include nonorthogonal predictor variables. These results will provide a practical basis for assessing the statistical significance of a regression developed by any empirical selection method.

Theoretical considerations often suggest the important independent variables and the functional form of the relationship. Models based on theoretical considerations are the exception rather than the rule in most practical managerial problems. As a result, the set of possible predictors may be quite large and the problem of selecting a "best" set becomes a difficult task. There are a number of articles in the literature, notably (11), (17) and (18), describing this problem and offering various criteria to be used to determine the variables to be included. Lindley (13) emphasizes that the selection criterion should be related to the intended use of the model. Hocking (11) gives a

description of these potential uses which include description, prediction, and control. It is generally recognized that a universally best criterion for selecting a set of predictor variables does not exist. It is not the purpose of this chapter to discuss the advantages and disadvantages of various selection procedures, but to determine a way of evaluating the statistical significance of the resulting model.

A commonly used selection procedure involves determining the adjusted multiple coefficient of determination, $R_a^2(k)$, for all $2^{p-1}$ subsets of the predictor variables where

$$R_a^2(k) = 1 - \frac{n(1-R^2(k))}{n-k}$$

and $R^2(k)$ is the coefficient of determination for the model with k predictor variables. The subset chosen is that with maximal $R_a^2(k)$. In fact, $R^2(k)$ plays a central role in almost all selection criteria. This value of $R^2(k)$, as previously shown, can be misleadingly large. How large must it be to be judged statistically significant? What makes the question hard to answer is the fact that the distribution of $R^2(k)$ for the selected model depends on the underlying relation among the variables as well as the selection criterion.

If p variables are being considered and all are included in the model, the classical F test is appropriate provided $p < n-1$. The test of

$$H_o: \rho^2_{y.12...p} = 0$$

is equivalent to testing

$$H_0: \rho_1 = \rho_2 = \ldots = \rho_p = 0$$

where $\rho_i$ is the simple correlation between the dependent variable, y, and the ith predictor, $x_i$. The null hypothesis is rejected if

$$F = \frac{R^2/(p)}{(1-R^2)/n-p-1)} > F(\alpha,p,n-p-1).$$

If $R^2$ is significant for the full model, a reduction in the number of predictors used is usually called for since a model with many independent variables is expensive to maintain, difficult to analyze and interpret, and almost always results in larger predictor variances (21) than a suitably selected submodel. The application of a selection procedure to obtain a "best" submodel may result in a submodel which is no longer statistically significant. Cramer (5) suggests that it is possible for the value of $R^2$ for the full model to be statistically significant while none of the regression coefficients have individually significant t values. In this situation, each predictor is making its own independent, albeit slight, contribution so that the total effect is statistically significant. It may not be possible to eliminate any variable or set of variables so as to maintain a significant $R^2$.

The F test cannot be used if the number of predictor variables is larger than n-2. Common sense and sound statistical practice require selection of a subset of predictor variables. When this selection is done empirically, as we have seen, $R^2$ (and hence F) becomes inflated so that the standard tests are invalid. Even though the distributional properties of $R^2$ under variable selection are hopelessly complex, an exact conditional test is derived which is valid for any variable selection technique.

Consider the hypothesis Ho: $\rho^2_{y \cdot 12 \ldots p} = 0$. Under Ho, the joint distribution of $\underline{X}$ and $\underline{Y}$ is invariant under permutations of $\underline{Y}$. Since there are n observations, there are $N = n!$ possible permutations of the y values. If the particular variable selection method being used is applied to each of these permutations, a set of N corresponding values of $R^2$ could, in principle, be generated. Let $R^2(i)$, $i = 1,2,\ldots,N$, be the ith smallest value in this set of N values, and let R denote the collection of order statistics so obtained. Let $R^2_*$ denote the value of $R^2$ associated with the unpermuted y values. $R^2_* \in R$ and, by invariance, $\text{Prob}(R^2_* > R^2(N-m)) = m/N$ for $1 \leq m \leq N$ so that the critical region $R^2_* > R^2(N-m)$ yields an exact level $\alpha = m/N$ test of Ho. This test is a special case of Fisher's randomization test. The power of this test will be discussed later in this chapter.

The following example will help illustrate the methodology. Consider the regression model

$$\underline{Y} = \begin{pmatrix} -2.10 \\ -.32 \\ .51 \\ -1.70 \end{pmatrix} = \beta_o \underline{1} + \underline{X}\underline{\beta} + \underline{\varepsilon} = \beta_o \underline{1} + \begin{pmatrix} 1.23 & -.04 \\ -.36 & .67 \\ -1.33 & .69 \\ .01 & .17 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \underline{\varepsilon}$$

where $(Y, X_1, X_2)'$ is distributed as $MVN(\underline{0}, \underline{I})$. There are $n = 4$ observations and $p = 2$ possible predictors. All possible regressions are calculated. The maximum adjusted $R^2$ criterion leads to the prediction equation

$$\hat{y} = -2.09 + 3.19\ x_2$$

with $R^2_* = .92$. Using this criterion to select a "best" subset for each of the 24 possible permutations of the y values yields the values of R shown in Figure 7. Note that $R^2_* = R^2(15)$; that is, $R^2_*$ is the 15th order statistic. If a value of $R^2$ is chosen at random from the set R.

$$\text{Prob}(R^2 > R^2_* \mid \text{Ho true}) = 9/24.$$

This result is compatible with Ho and gives little evidence to indicate that Ho is false.

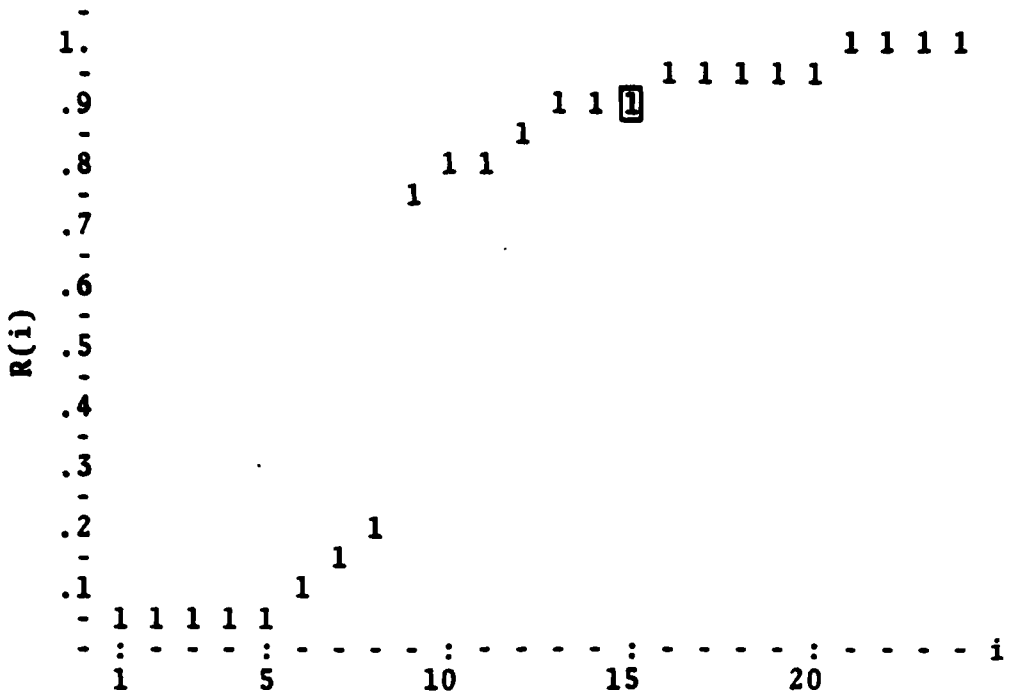



Figure 7: Plot of R(i) in the Set R

The statistical test illustrated in the above example is an exact test. Unfortunately, this test is not practical as a result of the large number of possible permutations that are required even for small values of n. For a sample of n = 25 observations, N would be over $10^{25}$. The calculations associated with this number of permutations make this approach computationally infeasible.

Since it is not practical to determine the entire set *R*, sampling schemes will be explored. Note that only the relative position of $R_*^2$ in *R* is needed in order to measure the probability of a Type I error for the conditional test. As a first approach to assessing the extremeness of $R_*^2$ relative to *R* we consider a nonparametric tolerance

interval argument (22). Based on a random sample of s permutations, where

$$s = \frac{\log(1-g)}{\log(1-d)},$$

it is known that

$$\text{Prob}(100(1-d)\% \text{ of the } R^2 \text{ in } \mathcal{R} < R^2(s)) > g$$

where $R^2(s)$ is the largest $R^2$ value in the sample. For example, with $g = .99$ and $d = .1$,

$$s = \frac{\log(.01)}{\log(.90)} = 44.$$

Thus, if 44 random permutations are obtained and their corresponding $R^2$ values calculated, then

$$\text{Prob}(90\% \text{ of the } R^2 \text{ in } \mathcal{R} < R^2(44)) > .99.$$

Comparing $R_*^2$ with $R^2(44)$ gives an indication of the relative position of $R_*^2$ in $\mathcal{R}$. This could provide the basis for a decision rule (reject Ho if $R_*^2 > R^2(44)$), but the significance level is only loosely related to the parameters g and d. For this example, the significance level would be approximately .1.

In order to obtain an exact test, this approach must be modified. Since extremely large values of $R_*^2$ relative to the set $\mathcal{R}$ provide evidence critical of the null hypothesis, the following decision rule is appealing: reject Ho if $R_*^2 > R^2(s)$ where $R^2(s)$ is the largest $R^2$ value in a

sample resulting from s random permutations. How large must s be in order that the test have a significance level of $\alpha$? If Ho is true, $R_*^2$ is just as likely to be any of the (s+1) observed $R^2$ values. That is,

$$\text{Prob}(R_*^2 > R^2(s) \mid \text{Ho is true}) = 1/(s+1).$$

A level $\alpha$ test is obtained by taking s permutations of the original y values where s is the smallest integer greater than or equal to $(1-\alpha)/\alpha$. For example, for $\alpha = .05$, $s = .95/.05 = 19$ permutations must be used.

The determination of the power that the permutation test will achieve against various alternatives is a difficult problem and remains unsolved. A simulation is employed to compare the power of the new test to that of the F test in some situations where the latter is valid. In these situations, the F test is optimal (1). But if the new test has comparable power, it will provide a alternative to the F test that is valid under a wider set of conditions, namely, when variable selection techniques are employed. In particular, random samples of size 30 are generated on the vector $(Y,X_1,X_2,\ldots,X_5)'$ which has a 6-dimensional multivariate normal distribution with mean vector $\underline{0}$ and covariance matrix $\underline{C}$. The data is analyzed by fitting the full model

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon,$$

and no variable selection technique is used. One hundred such samples are generated and analyzed for each distribution. For each test, the fraction of these samples which resulted in $H_o$ being rejected gives an indication of that test's power.

The covariance matrix $\underset{\sim}{C}$ is of the form

$$\underset{\sim}{C} = \left(\begin{array}{c|ccc} 1 & r_1 & r_2 & \cdots & r_5 \\ \hline r_1 & & & & \\ r_2 & & & & \\ \vdots & & \underset{\sim}{D} & & \\ r_5 & & & & \end{array}\right)$$

The values of $r_1, r_2, \ldots, r_5$ and $\underset{\sim}{D}$ are chosen to give prespecified values for the theoretical coefficient of determination, $\rho^2$, and to allow for various covariance structures such as nonorthogonal predictor variables. It is known (1) that in these situations the power of the F test depends on the covariance structure only through the value of $\rho^2$. The purpose of this simulation is to make a comparative study of the powers of the permutation test and F test for the following special covariance structures.

Case 1. $\underset{\sim}{D} = \underset{\sim}{I}$ and $r_i = r_1^*$ for $i = 1,2,\ldots,5$.

Case 2.

$$\underset{\sim}{D} = \underset{\sim}{D}_1 = \begin{pmatrix} 1 & .9 & .9 & \cdots & .9 \\ & 1 & .9 & \cdots & .9 \\ & & \ddots & & \vdots \\ & & & 1 & .9 \\ & & & & 1 \end{pmatrix}$$

and $r_i = r_2^*$ for $i = 1,2,\ldots,5$.

Case 3. $\underset{\sim}{D} = \underset{\sim}{D}_1$ and $r_1 = r_2 = r_3^*$ and $r_i = 0$ for $i = 3,4,5$. The values of $r_1^*$, $r_2^*$, and $r_3^*$ are chosen to give the desired $\rho^2$ values.

The results of this simulation appear in Table III. The fraction rejected by both tests is given for various $\rho^2$ values for each of the three cases. The theoretical power of the F test also appears in the table. While these theoretical values are listed, the actual fractions rejected by the F test are included to give a better basis for comparing the corresponding fraction rejected by the permutation test based on the same data.

TABLE III.

Fraction Rejected at .05 Significance Level

| $\rho^2$ | Actual Power of F test | Case 1 | | Case 2 | | Case 3 | |
|---|---|---|---|---|---|---|---|
| | | Perm. | F | Perm. | F | Perm. | F |
| 0 | .050 | .06 | .06 | .05 | .07 | .05 | .04 |
| .1 | .196 | .15 | .18 | .20 | .19 | .21 | .23 |
| .2 | .418 | .36 | .44 | .33 | .39 | .39 | .42 |
| .3 | .657 | .55 | .67 | .54 | .61 | .60 | .69 |
| .4 | .847 | .78 | .85 | .75 | .83 | .80 | .87 |
| .5 | .954 | .90 | .95 | .91 | .94 | .89 | .96 |
| .6 | .992 | .97 | .98 | 1.00 | .99 | 1.00 | 1.00 |
| .7 | .999 | 1.00 | .99 | 1.00 | 1.00 | 1.00 | 1.00 |
| .8 | .999+ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| .9 | .999+ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

While the power of the permutation test cannot be expected to match that of the F test, these figures offer evidence that it performs surprisingly well. For fixed $\rho^2$ values, the covariance structure appears to have no significant effect on the power of the permutation test although that conjecture remains an open question. To obtain more numerical insight into that question, a larger scale simulation was performed for $\rho^2 = .4$ by running 6 independent replications of size 100 for each covariance structure.

TABLE IV

Fraction Rejected at .05 Significance Level for $\rho^2 = .4$ - Additional Data

| Replication | 1 | 2 | 3 | 4 | 5 | 6 | Mean | Stand. dev. |
|---|---|---|---|---|---|---|---|---|
| Case 1 | .71 | .82 | .77 | .78 | .76 | .84 | .78 | .046 |
| Case 2 | .70 | .79 | .82 | .71 | .81 | .74 | .76 | .051 |
| Case 3 | .81 | .78 | .80 | .82 | .70 | .74 | .775 | .046 |

These results certainly strengthen the credibility of the conjecture that the power of the permutation test depends on the covariance structure only through the value of $\rho^2$ when no variable selection technique is used.

These results give some indication of the relative performance of the permutation and F tests in situations where the F test is valid. While the exact power function of the F test is known (1), the mathematically untractable power function of the permutation test necessitates this Monte-Carlo approach. In Chapter V, the application of this test to problems of interest to management science will be investigated.

# CHAPTER IV

## POWER TRANSFORMATIONS OF BIVARIATE SAMPLES

In the analysis of data it is often necessary to use a power transformation to model the relationship between two variables. An appropriate transformation may be suggested by economic theories, physical properties, or other such underlying considerations of the relation being studied. On the other hand, there may be an absence of any such firm theoretical or even historical indications. Upon inspection of the scatter diagram, it may be obvious that a linear relationship between the two variables is not appropriate. For these situations, Mosteller and Tukey (19) suggest considering a re-expression of one or both of the variables so that the resulting relationship is more nearly linear. They suggest a re-expression of the form $(y+c)^p$ where c and p are constants. According to Mosteller and Tukey, the value of c is often zero and the most commonly used powers are $p = 1/2$, $p = -1$, and $p = 1/3$ in descending frequency of use.

To aid practitioners in the selection of possible p values, they offer a rule of thumb called the "bulging rule." Using the scatter diagram in accordance with this rule indicates what values of p should be considered. For

the original y values, p is equal to 1. From this value of p, the fundamental rule is to move on the "ladder" of possible values of p in the direction in which the bulge of the scatter plot points. Figure 8 illustrates how to use this ladder of powers to aid in the re-expression of y for four kinds of bulging. If the scatter plot resembles curve (a), movements up the latter and values of p larger than 1 should be considered. The "bulging rule" may also be used to indicate appropriate values for power transformations of x as illustrated.

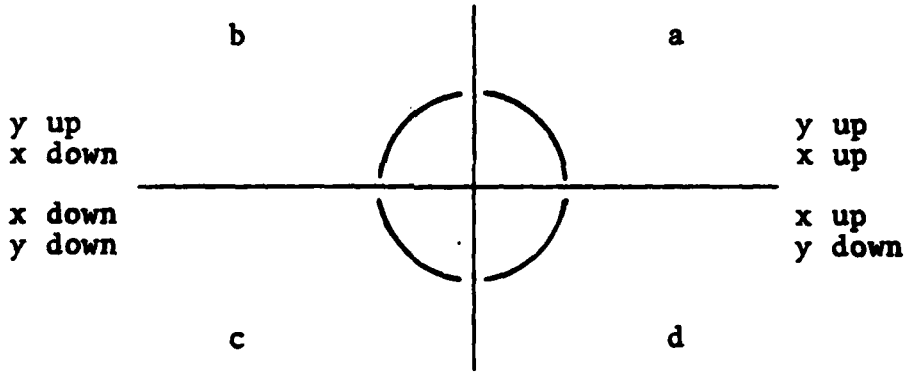



Figure 8: The Bulging Rule

The purpose of this chapter is to report the results of a study of the influence of this power transformation on the sample linear correlation between the two variables when the values of c and p are empirically determined in such a way as to maximize the sample linear correlation R. Maximizing R by empirically determining c and p does not coincide with Tukey's notion of "straightening out" the data. The focus

of this investigation is the possibility of artificially high values for R when the true linear correlation is zero. The motivation for this study was the question, "Is it possible to significantly increase the sample linear correlation between two variables by considering power transformations of the dependent variable when the variables are, in fact, independent?"

A simulation is performed in an attempt to answer this question. Samples of size 10 are generated from a bivariate normal population with mean vector $(10, 10)'$ and covariance matrix $\underset{\sim}{I}$. A mean of 10 and a standard deviation of 1 are used to insure positive values for y since, according to Mosteller and Tukey, y should represent an amount or count if the re-expression $(y+c)^p$ is to be used. As a result of this covariance structure, the theoretical linear correlation is zero. Let $R^*(c,p)$ denote the sample correlation between $(y+c)^p$ and x. Note that $R^*(c,1) = R$ for all c. The values of c and p are determined via a two-dimensional optimizing program in such a way as to maximize the value of $R^*(c,p)$. Let $R^*$ denote this maximum value. A number of such samples are considered with R and $R^*$ values calculated for each sample. A scatter plot of these pairs is given in Figure 9. For certain samples, the sample linear correlation is substantially increased.

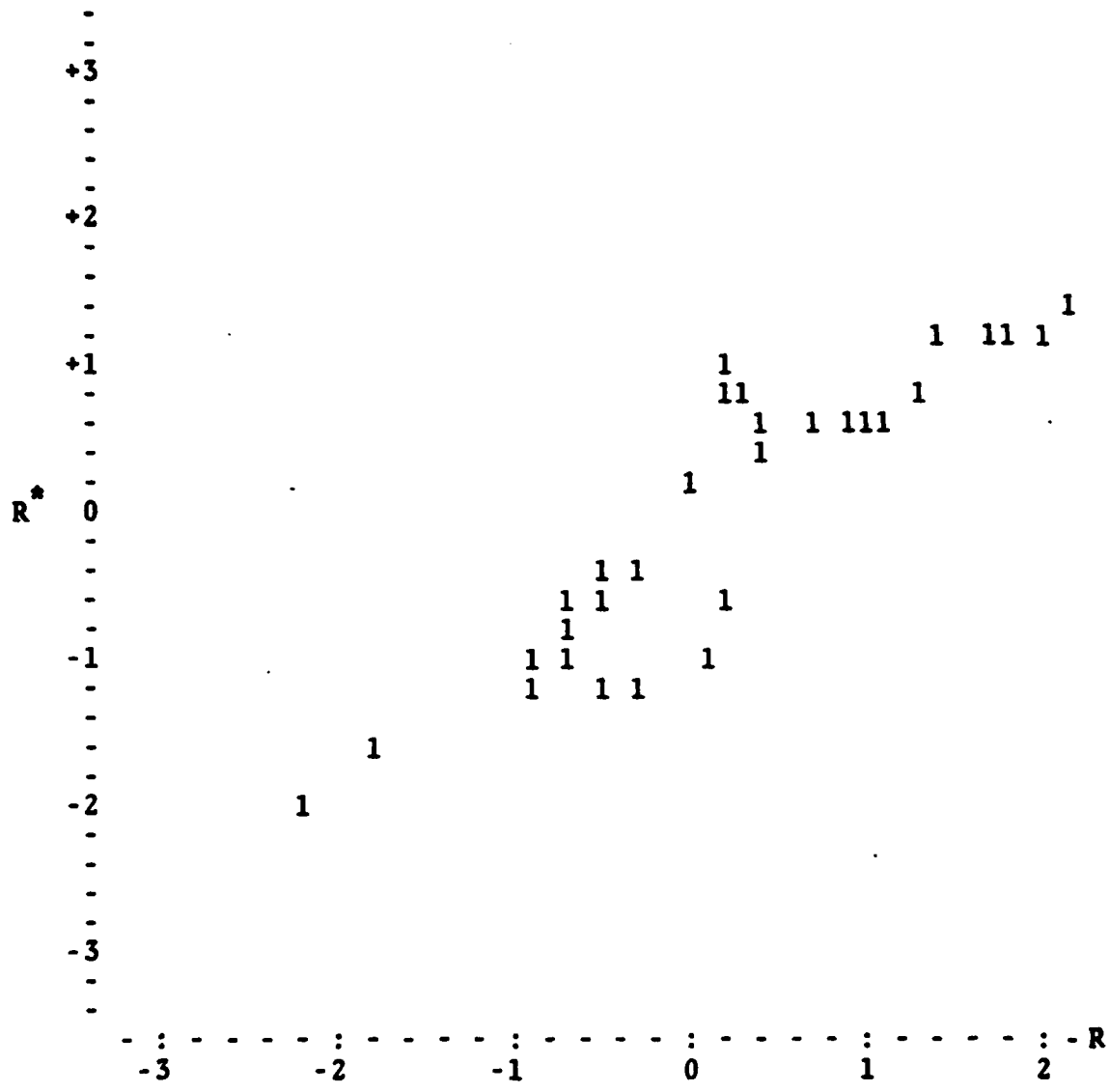



Figure 9: Scatter Plot of 30 Standardized Values of $R^*$ vs R

A simple linear regression of $R^*$ on R is performed in order to investigate their relationship further. The resulting estimate of the slope is 1.27 which is found to be significantly greater than 1 at the .05 level. Figure 10 gives a plot of the absolute, standardized values of the residuals which deserves some attention. Note that the

residuals reflect the fact that for extreme values of R, values near -1 or 1, the amount of inflation is not as severe as in cases where R is originally small. Thus, if the initial correlation is small in absolute value, there is a greater potential for the power transformation to result in an inflated value of $R^*$. This potential will be investigated in more detail below.

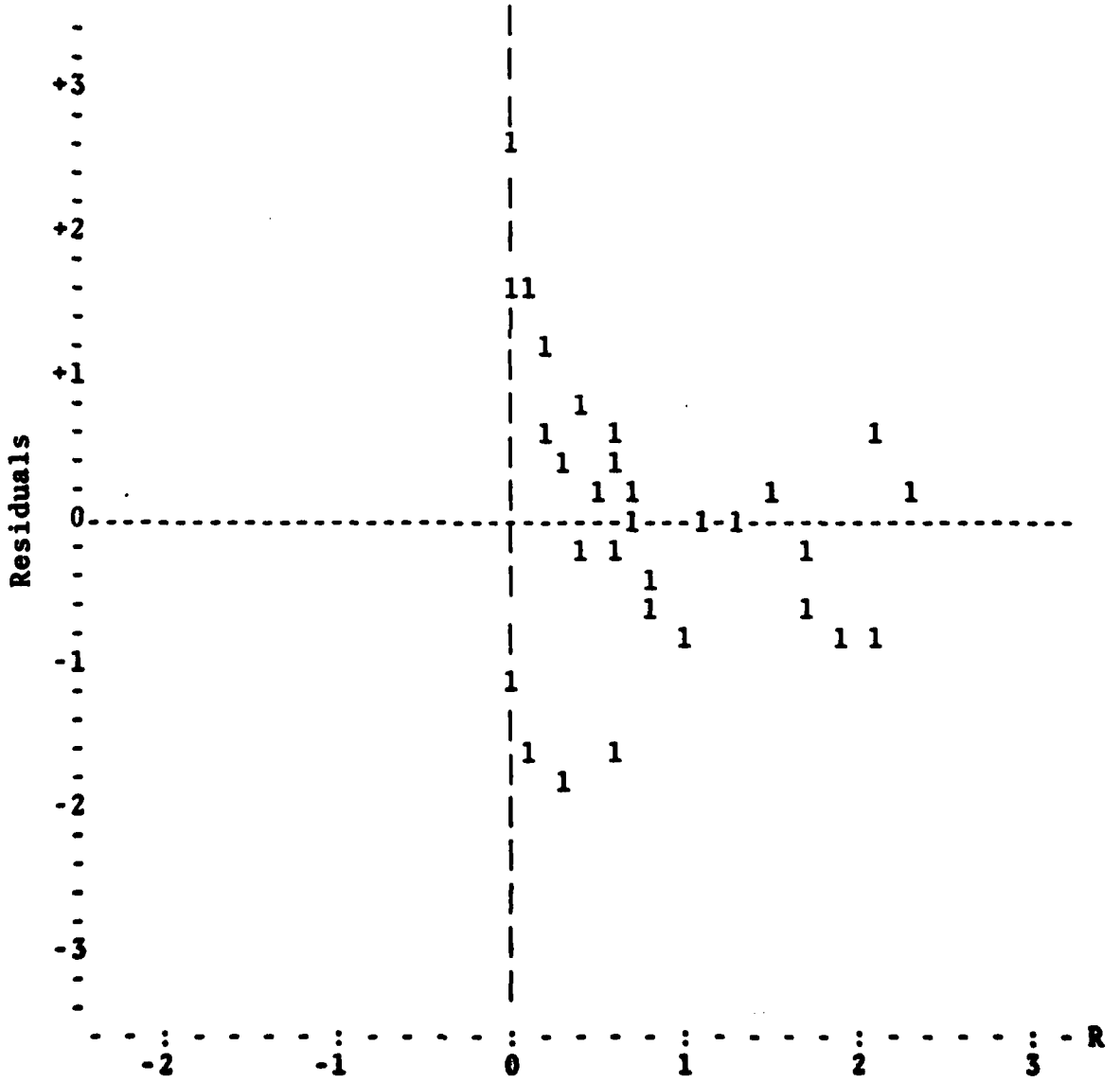



**Figure 10: Scatter Plot of 30 Absolute Standardized Values of Residuals vs R**

Examination of the simulation results reveals the fact that the optimal transformations seem to cluster into two categories. For each of these categories, the optimal value of c tends to be the negative of the smallest y observation. The optimal p values are either in the range 1.5-4.5 or the search algorithm fails to identify an optimal p value. In the latter case, $R^*(-y(1),p)$ increases as p is decreased toward zero. Examination of scatter plots reveals the reason for this phenomenon. It is related to the so-called "lollipop effect" whose name will become clear shortly.

Recall that

$$R = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left\{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}^{\frac{1}{2}}\left\{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2\right\}^{\frac{1}{2}}} = \frac{s(x,y)}{s(x)s(y)} .$$

Obviously, if there is no linear relationship between x and y, s(x,y) would be expected to be close to zero. Since only transformations of y are being considered, the x values and, thus, s(x) are fixed. Therefore, a transformation of y yields a variable $y^*$ that will be more linearly correlated with x than y is if the resulting ratio $s(x,y^*)/s(y^*)$ is larger than s(x,y)/s(y). When can such a possible transformation be expected to exist? The answer lies in the analysis of the scatter plot of the (x,y) pairs. If an extreme value of y is associated with an extreme value of x, then it is possible to use a transformation of the form $y^* = (y+c)^p$

to substantially increase the sample linear correlation as shall be seen with the help of an example.

Table V gives a set of x and y pairs resulting from a typical simulation run. Note that the smallest value of $y = 8.79$ is paired with $x = 8.89$, the third smallest x. Also, the deviation of this x from the mean, $\bar{x}$, is -.96, the third largest in absolute value. Some of the initial statistics are $R = .11$, $s(x,y) = .109$, $s(y) = 1.25$ and $\bar{y} = 10.51$. A transformation with parameter values as described for the second type of optimality is found using the search algorithm. That is, the smallest y value (8.79) is subtracted from each of the y observations, and the resulting differences raised to the power $p = 0$. These values are given in Table V. Note that $s(y^*) = .317$ is much smaller than $s(y) = 1.25$. Thus, the influence of the transformation on the value of $s(x,y^*)$ will determine if the sample linear correlation is increased. Note that $s(x,y^*)$ is a weighted sum of deviations of x values about their mean. These weights are the differences $(y_i^* - \bar{y}^*)$. As a result of the transformation, these deviations are small except for $(y_1^* - \bar{y}^*) = -.90$. The transformation reduces the variability of the dependent variable and associates with a large x deviation the largest $y^*$ deviation, which yields $s(x,y^*) = .106$. As a result of this transformation, R is increased from .11 to $R^* = .41$.

TABLE V.

Values from Simulation Run

| y | x | $y^*$ |
|---|---|---|
| 8.79 | 8.89 | 0 |
| 8.93 | 9.80 | 1 |
| 11.06 | 10.77 | 1 |
| 12.21 | 10.33 | 1 |
| 10.68 | 10.51 | 1 |
| 10.96 | 9.70 | 1 |
| 10.15 | 8.58 | 1 |
| 10.44 | 10.67 | 1 |
| 12.46 | 8.86 | 1 |
| 9.37 | 10.36 | 1 |

$\bar{y} = 10.51$ $\bar{x} = 9.85$ $\bar{y}^* = .90$

$s(y) = 1.25$ $s(x) = .82$ $s(y^*) = .317$

$s(x,y) = .109$ $s(x,y^*) = .106$

$R = .11$ $R^* = .41$

This example illustrates the "lollipop effect." The name is a result of the appearance of the scatter diagram of the $(x,y^*)$ pairs (see Figure 11). The optimal transformation isolates one point while grouping the remaining points giving the data set a lollipop appearance.

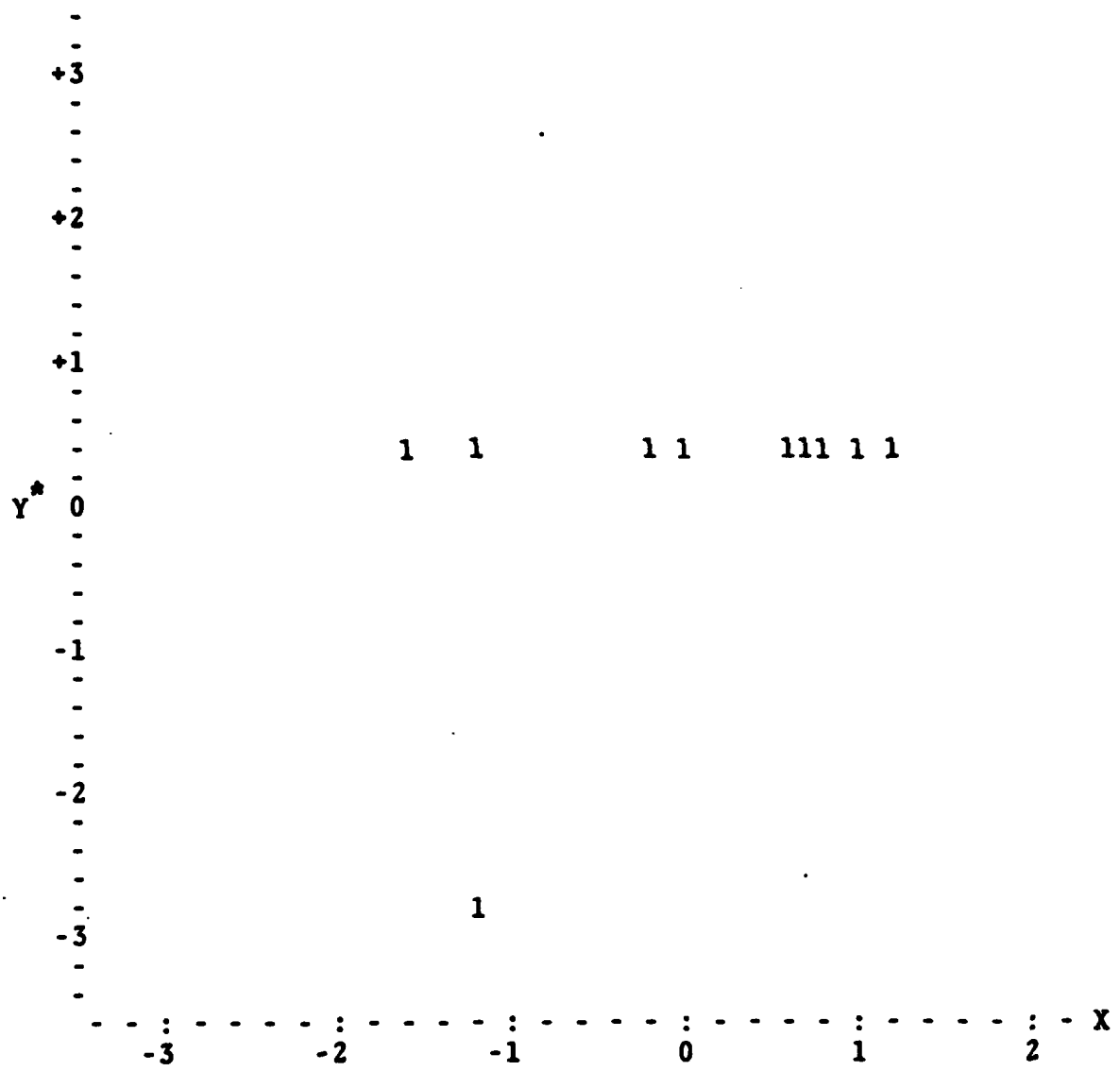



Figure 11: Scatter Plot of 10 Standardized Values of $Y^*$ vs. X

It has been shown that it is possible to inflate the value of R using an empirically-determined power transformation when the two variables are actually independent. However, this "inflation" should not go undetected. A residual plot, such as Figure 12 for the above example, indicates

this lollipop effect. Upon inspection of such residuals, the experienced data analyst should not fail to realize the reason for this anomaly and reject the method summarily.

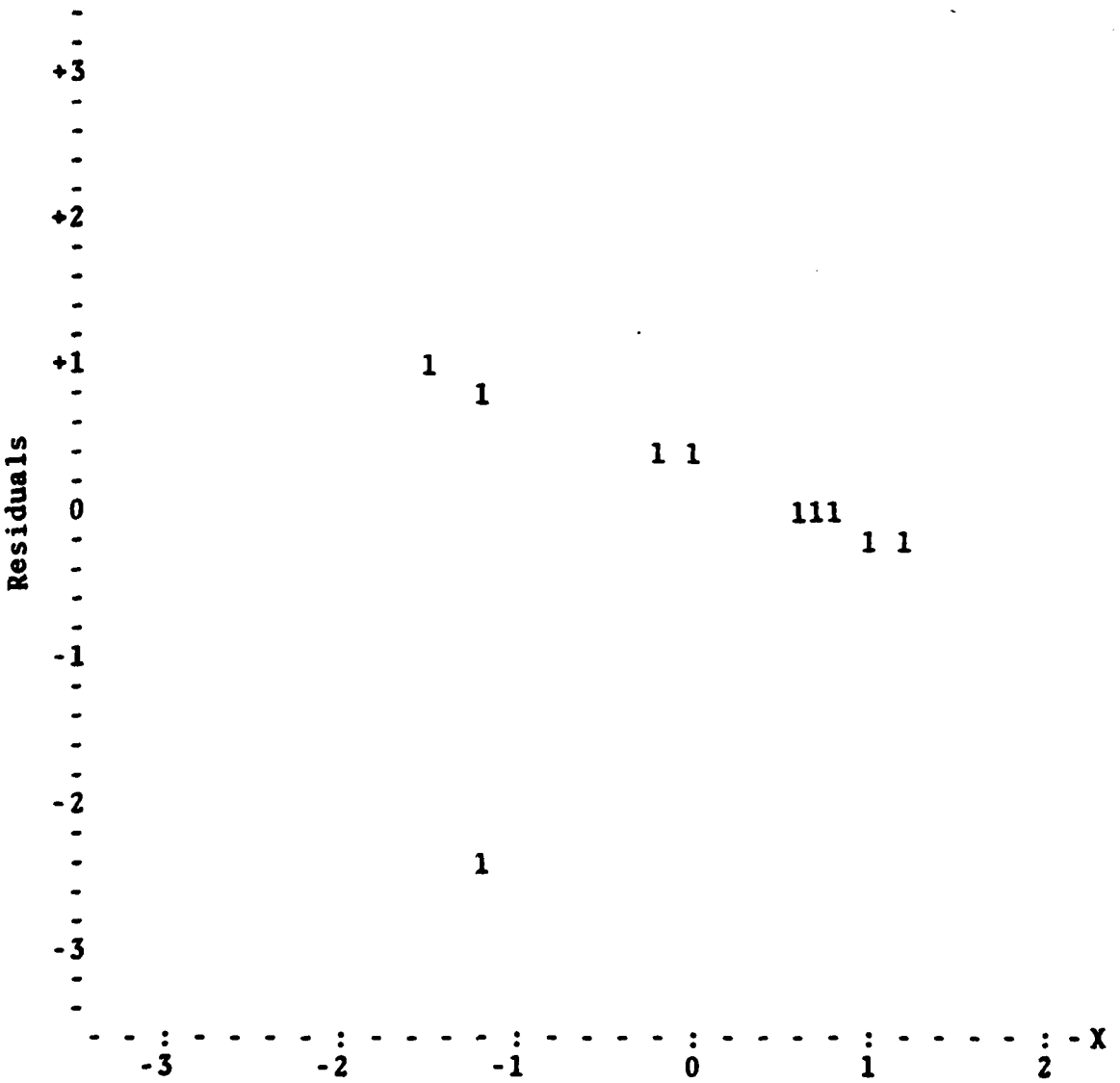



Figure 12: Scatter Plot of 10 Standardized Values of Residuals vs. X

For some samples, it has been shown that the "lollipop effect" creates a substantially increased value of the sample linear correlation. Even samples from correlated, bivariate populations may have the potential for this phenomenon. The following theorem gives insight into this potential for any bivariate sample.

Theorem: For a bivariate sample of size n, $(y_1,x_1)'$, $(y_2,x_2)',\ldots,(y_n,x_n)'$, let $y_j$ be the smallest y value in the sample and let y be transformed to $y^*$ so that $y_i^* = 1$ for $i = 1,2,\ldots,n$, $i \neq j$ and $y_j^* = 0$. Then the value of the simple linear correlation, $R^*$, between $y^*$ and x is given by

$$R^* = \frac{\sqrt{n}(\bar{x}-x_j)}{(n-1)s(x)}$$

and the estimate of the parameters in the regression of $y^*$ on x, given by

$$y^* = b_0 + b_1x + e,$$

are

$$\hat{b}_1 = \frac{\bar{x}-x_j}{(n-1)(s(x))^2} \quad \text{and}$$

$$\hat{b}_0 = \frac{n-1}{n} - \hat{b}_1\bar{x}.$$

Proof: Obviously, $\bar{y}^* = (n-1)/n$. Thus

$$[s(y)]^2 = \frac{1}{n-1}\left(\sum_{\substack{i=1 \\ i\neq j}}^{n}\left(1 - \frac{(n-1)}{n}\right)^2 + \left(0 - \frac{(n-1)}{n}\right)^2\right) = \frac{1}{n}.$$

Therefore,

$$R^* = \frac{\frac{1}{n-1}\left(\sum_{\substack{i=1 \\ i\neq j}}^{n}(x_i-\bar{x})\left(1 - \frac{(n-1)}{n}\right) + (x_j-\bar{x})\left(0 - \frac{(n-1)}{n}\right)\right.}{s(x)\sqrt{1/n}}$$

$$= \frac{\sqrt{n}}{(n-1)}\,\frac{\bar{x}-x_j}{s(x)}.$$

Furthermore,

$$\hat{b}_1 = R^*\,\frac{s(y^*)}{s(x)} = \frac{(\bar{x}-x_j)}{(n-1)(s(x))^2}$$

and

$$\hat{b}_0 = \bar{y}^* - \hat{b}_1\bar{x} = (n-1)/n - \hat{b}_1\bar{x}. \qquad \text{Q.E.D.}$$

If the sample deviation, s(x), is "small" relative to the deviation from the mean of the x value corresponding to the smallest y observation, then the resulting $R^*$ may be substantially inflated. Two examples will help illustrate this point.

In the first example, a sample of size 10 is taken from a bivariate normal population with mean vector $(10, 10)'$ and covariance matrix

$$\begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}.$$

A scatter plot of these points appears in Figure 13. The theoretical correlation is .5, and the sample correlation is .66. Note that the smallest value of y which is 8.54 is associated with the smallest x value which is 7.04. Thus, an inflated value of $R^*$ is expected. Using the above theorem, we have

$$R^* = \frac{\sqrt{n}\ (\bar{x}-x_j)}{(n-1)s(x)} = \frac{\sqrt{10}\ (10-7.04)}{(9)(1.26)} = .84.$$

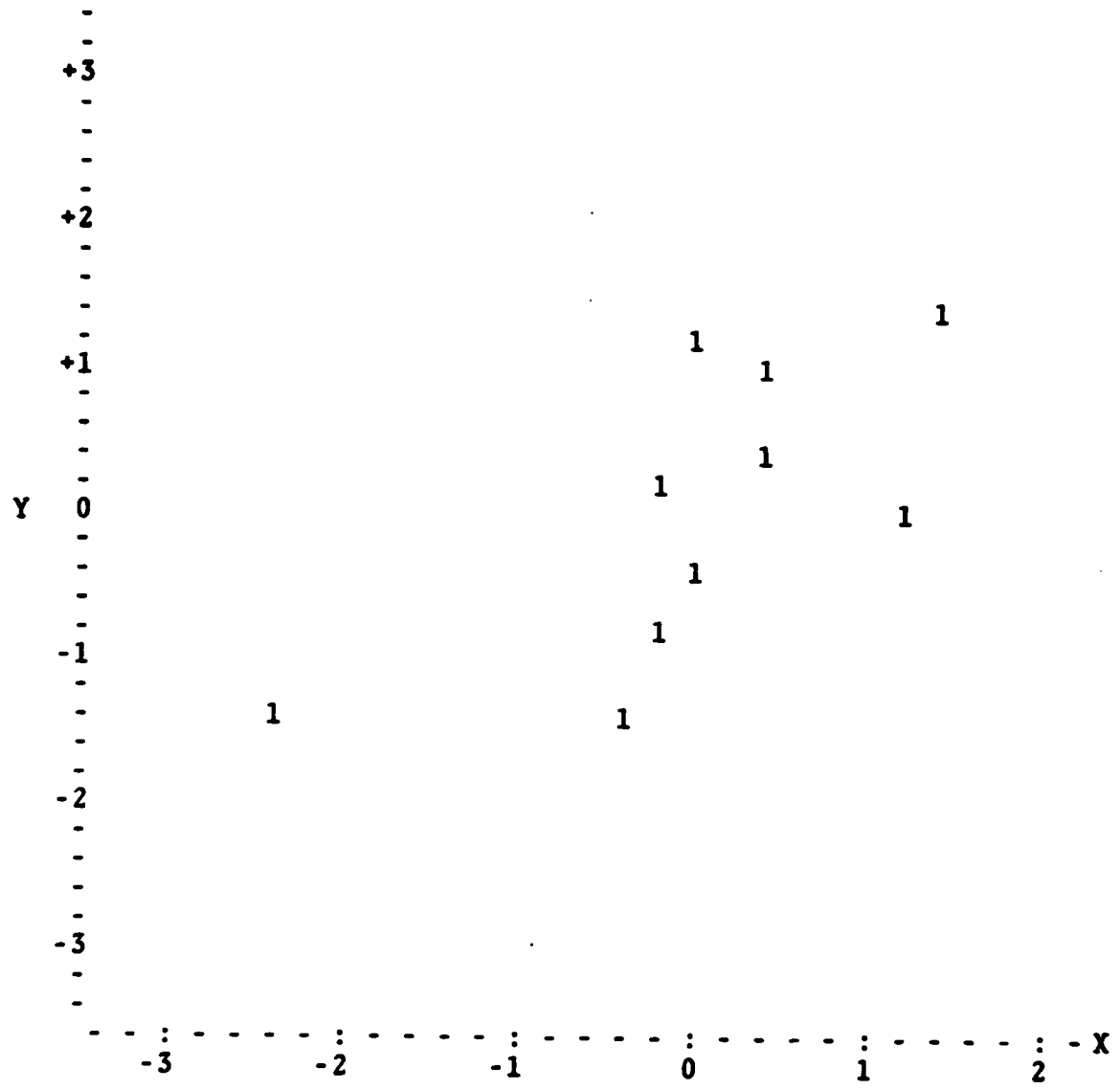



Figure 13: Scatter Plot of 10 Standardized Values
Y vs. X

Figure 14 is the scatter plot of 10 observations from a bivariate normal population with mean vector $(10, 10)'$ and covariance matrix

$$\begin{pmatrix} 1 & .8 \\ .8 & 1 \end{pmatrix}.$$

The theoretical correlation is .8, and the sample correlation is .82. Note that the smallest y, 9.36, is paired with an x, 9.67, which is near its sample mean of 10.18. Thus, the above transformation might result in a small value of $R^*$ as seen by

$$R^* = \frac{\sqrt{10}\,(10.18 - 9.67)}{(9)(.77)} = .34.$$

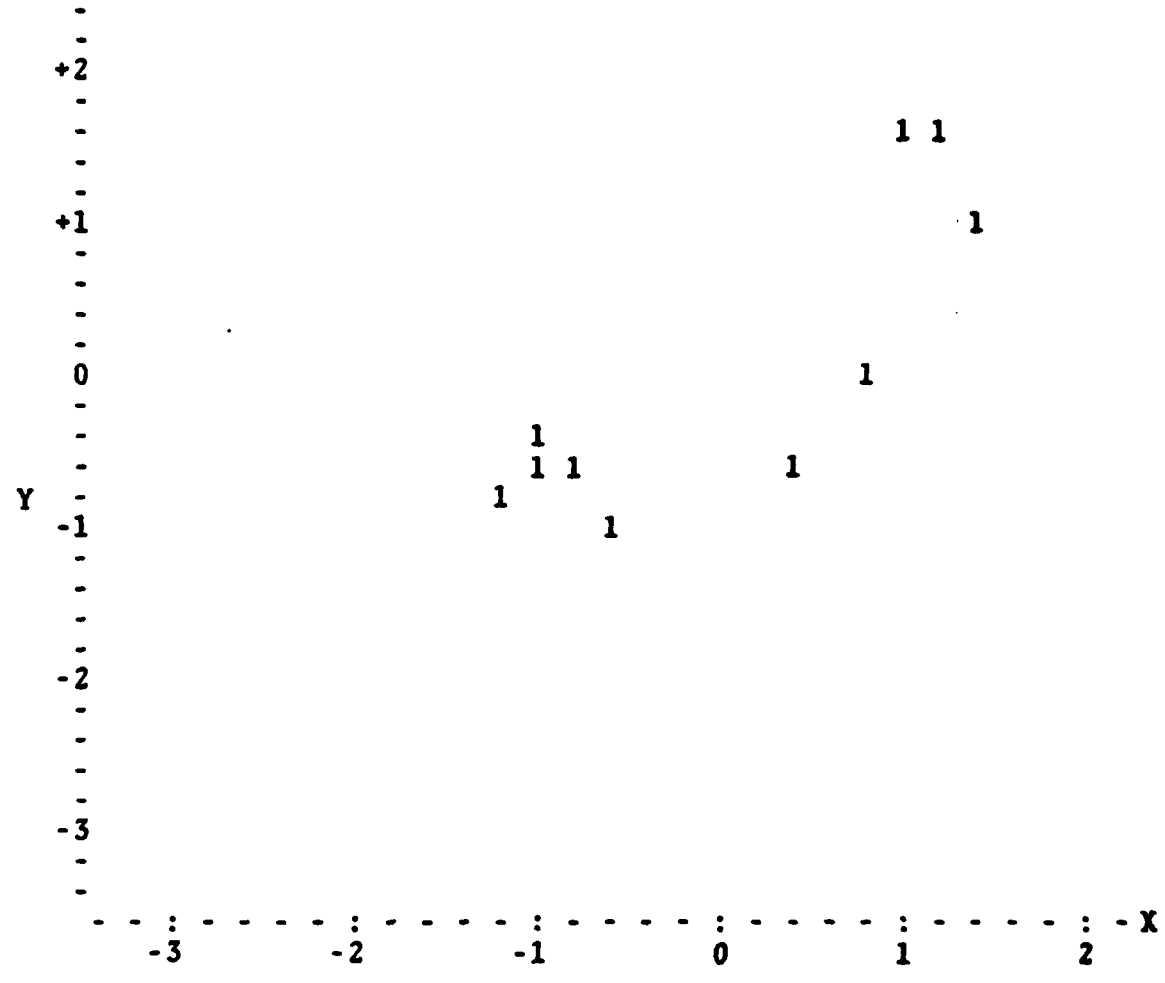



Figure 14: Scatter Plot of 10 Standardized Values
Y vs. X

The use of empirically-determined power transformations may lead to inflated values of the sample linear correlation. The potential for this phenomenon has been demonstrated for independent and correlated bivariate samples. However, the fact that the value is inflated should not go undetected. An inspection of the resulting residuals should aid the data analyst in spotting this "lollipop effect".

# CHAPTER V

## APPLICATIONS OF PERMUTATION TEST

In this chapter the applicability of the statistical procedures developed previously to problems of parametric cost estimation is illustrated. Parametric cost estimation is a management tool used to aid in the prediction of the cost of a proposed system. It involves predicting the cost (dependent variable) of a system by means of explanatory (independent) variables such as system characteristics or performance requirements. This procedure is based on the premise that the cost of a system is related in a quantifiable way to the system's physical and performance characteristics (14). The expression of this quantifiable relationship is in the form of an estimating equation derived through statistical regression analysis of historical cost data on systems which are, more-or-less, analogous to the proposed system. Since parametric cost estimates can be developed during the concept formulation stage of the acquisition process before engineering plans are finalized, these estimates can be used by management to (14):

1. Identify possible cost/performance tradeoffs in the design effort.
2. Provide a basis for cost/effectiveness review of performance specifications.
3. Provide information useful in the ranking of competing alternatives.

4. Suggest a need for investigating new alternatives.

In particular, examples of parametric cost estimation for Navy weapon systems will be considered. Cost overruns have been prevalent in the acquisition of new weapon systems making cost estimation a very important problem for all components of the Department of Defense. These overruns result in very difficult budget decisions and a decrease in the Congress's confidence in the managerial ability of military leaders. For fiscal year 1971, the Navy experienced a cost growth of $19 billion on 24 weapon systems; 15% of this cost growth was attributed to poor initial cost estimates (14). Historically, the Navy has used industrial engineering techniques to develop estimates of the cost of a proposed system. These techniques required detailed studies of the operations and materials required to produce the new system. Although a great deal of time and effort is required to produce these estimates, there is considerable uncertainty remaining as evidenced by the overruns mentioned above. In addition, slight design changes can vitiate the estimate and neccessitate a complete restudy. To help improve such performance, the Department of Defense has issued directives to all branches of the service to employ independent parametric cost estimation. Publications such as (14) have appeared which give step by step methodology for the development of a parametric cost estimate.

Regression problems faced by costing and pricing analysts in these situations are inherently difficult for two fundamental reasons (20):

1. The number of observations is usually small compared with the number of system characteristics which are candidate components of the regression equation.

2. The available data is not produced by employing an efficient experimental design, but by what Box (4) calls "unplanned happenings."

Under these circumstances, it has been shown that the use of variable selection techniques may result in regression equations which yield inflated $R^2$ values whose statistical significance cannot be tested using the F test.

One approach for the development of a parametric cost estimate involves breaking the system up into component subsystems and using a separate model to estimate the cost of each component. This process, called disaggregation (14), will generally result in better subsystem cost estimates, and if these estimates are independent, a combined estimate of system cost can be obtained in the obvious way. For example, a cost estimate may be desired for the construction of a new submarine under consideration. A possible componet subsystem would be its sonar system. A cost estimate of this subsystem might be based on a model with such candidate predictors as weight and volume of the internal electronics, number of hydrophone amplifiers, power output, sensitivity, the year that the sonar system became fleet operational, etc. Total system cost is then estimated by

reaggregating subsystem estimates. The determination of a confidence interval for the total system cost is a difficult problem and remains unsolved. The difficulty is a result of the lack of understanding of the effect of interactions among the subsystems on the factors influencing total cost. The development of cost-estimating models for missile subsystems will be explored via data obtained from the Naval Weapons Center at China Lake, California. The data has been sanitized for security reasons without destroying the relationships between variables.

Table VI presents historical data on the cost and relevant performance characteristics of a certain type of system which we shall designate Subsystem A. Presumably, values of $X_1, X_2, \ldots, X_7$ of a proposed system will be substituted into the prediction equation developed for the data in Table VI in order to produce a cost estimate of the proposed system. As is typical for parametric costing problems, the number of observations available is not large compared to the number of candidate predictors. Here, there are 8 observations on the cost and 7 system characteristics. With the information provided in this data, we want to determine the performance characteristics which provide a model that will best estimate the cost of the proposed subsystem.

TABLE VI.

Cost Data for Subsystem A

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | Y |
|---|---|---|---|---|---|---|---|
| 1.09 | 2.06 | 0.41 | 2.48 | 1.08 | 0.00 | 0.00 | 0.00 |
| 1.09 | 2.06 | 0.41 | 2.48 | 2.17 | 0.03 | 0.00 | 0.02 |
| 1.09 | 2.06 | 0.41 | 2.48 | 2.17 | 0.03 | 0.00 | 0.04 |
| 0.00 | 0.21 | 0.00 | 0.00 | 0.00 | 2.12 | 0.56 | 0.78 |
| 0.00 | 0.62 | 0.12 | 0.19 | 0.00 | 2.22 | 0.64 | 0.85 |
| 2.38 | 0.00 | 1.39 | 1.24 | 0.54 | 1.89 | 2.08 | 2.47 |
| 2.38 | 0.00 | 1.39 | 1.24 | 2.17 | 1.84 | 2.08 | 1.96 |
| 2.38 | 0.00 | 2.99 | 1.24 | 2.17 | 0.91 | 2.17 | 1.94 |

A stepwise regression algorithm is applied to the data yielding:

1. the best single-variable model

$$\hat{y} = .06 + .98\, x_7, \qquad (1)$$

2. the best 2-variable model

$$\hat{y} = .20 - .12\, x_5 + .99\, x_7. \qquad (2)$$

The $R^2$ associated with model (1) is .964 and that with model (2) is .978. Thus, the data analyst might consider using the single-variable model to obtain a cost estimate since, as mentioned in a previous chapter, the variance of prediction cannot be reduced by adding variables to the regression equation. The standard F test applied to this model yields a highly significant

$$F = \frac{R^2(n-p-1)}{(1-R^2)p} = 159.57.$$

Having shown that the use of variable selection tends to inflate the value of the F statistic, we consider the permutation test. Since a significance level of .05 is desired, 19 random permutations must be used. For each permutation, the stepwise algorithm is used to determine the best single-variable model. The $R^2$ value associated with this model is saved. Recall that the rejection rule is to reject Ho: $\rho^2 = 0$ if $R_*^2$, which is .964, exceeds $R^2(19)$ where $R^2(19)$ is the largest $R^2$ observed in the sample of permutations. Figure 15 gives a stem and leaf display of the 20 $R^2$ values.

Note that the largest sample value of $R^2$ is .897. Thus, $R_*^2 > R^2(19)$, Ho is rejected, and it is concluded that the single-variable model is significant at the .05 level. A cost estimate for the proposed system is obtained by evaluating this single-variable model at the value $X_7$ of the proposed system.

| | |
|---|---|
| 1.0 | |
| 0.9 | 64 (=$R_*^2$) |
| 0.8 | 97 |
| 0.7 | |
| 0.6 | 13,92 |
| 0.5 | 13,21 |
| 0.4 | 12,57,64 |
| 0.3 | 09,43 |
| 0.2 | 22,26,27,35 |
| 0.1 | 42,55,66 |
| 0.0 | 25,68 |

Figure 15: Stem and Leaf of $R^2$ Values for Subsystem A

Historical information for systems similar to a proposed system designated as Subsystem B appears in Table VII. Six observations are supplied on the cost of the system and 7 of its operating characteristics. Again a step-wise algorithm is applied to the data, and it yields the following models:

1. $\hat{y} = -.23 + .88\ x_1$, (3)

2. $\hat{y} = -1.19 + .96\ x_1 + .46\ x_7$. (4)

The $R^2$ values for models (3) and (4) are .768 and .978, respectively. The two-variable model appears to do the better job, but both will be analyzed.

TABLE VII.

Cost Data for Subsystem B

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | Y |
|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 0.72 | 0.55 | 1.95 | 2.94 | 2.04 | 0.00 |
| 0.79 | 1.57 | 0.72 | 0.00 | 0.24 | 1.56 | 2.04 | 0.42 |
| 0.79 | 1.57 | 0.72 | 0.00 | 0.24 | 1.56 | 2.04 | 0.37 |
| 2.91 | 1.57 | 2.40 | 0.00 | 2.15 | 1.10 | 2.26 | 2.74 |
| 0.79 | 2.32 | 2.40 | 0.34 | 0.00 | 0.00 | 2.99 | 0.85 |
| 1.58 | 0.17 | 0.00 | 2.56 | 0.04 | 0.64 | 0.00 | 0.28 |

The F test yields: for (3), F = 13.33 > F(.05,1,4) = 7.71 and for (4) F = 66.93 > F(.05,2,3) = 9.55. Therefore, both models appear to be significant at the .05 level. Once again the permutation test is applied using 19 random permutations of the cost values. Figures 16 and 17 present stem and leaf histograms for the $R^2$ values associated with the best one- and two-variable models, respectively.

| | |
|---|---|
| 1.0 | |
| 0.9 | 45 |
| 0.8 | 02,22,36 |
| 0.7 | 28,28,68 (=$R^2_*$),88 |
| 0.6 | 62,66,70 |
| 0.5 | 24 |
| 0.4 | 82,97 |
| 0.3 | |
| 0.2 | |
| 0.1 | 00,44,78 |
| 0.0 | 70,79,98 |

Figure 16: Stem and Leaf of $R^2$ Values for One-variable Models

| | |
|---|---|
| 1.0 | |
| 0.9 | 62,72,78(=$R^2_*$),78,89 |
| 0.8 | 31,31,32,44,64,78 |
| 0.7 | |
| 0.6 | 20,28,90 |
| 0.5 | |
| 0.4 | 09 |
| 0.3 | 13 |
| 0.2 | 10,14 |
| 0.1 | 60,87 |
| 0.0 | |

Figure 17: Stem and Leaf of $R^2$ Values for Two-variable Models

First of all, consider the results for the one-variable model. The largest value of $R^2$ in the sample is .945, which prohibits the rejection of $H_o$. Also, note that the value of $R^2$ associated with the unpermuted data, .768, is surpassed by a number of other sample $R^2$ values. This tends to give more evidence that the one-variable model given by (3) is not statistically significant. For the two-variable model, the largest sample value of $R^2$ is .989. Once again Ho cannot be rejected. The two-variable model appears not to be statistically significant. Based upon these results, the analyst would conclude that none of the proposed models provide a statistically significant fit for the cost of this subsystem.

In general parametric cost estimation, a researcher should not blindly trust the regression equation resulting for his analysis. To measure the "goodness of fit", the

analyst can use such statistics as $R^2$ and F. However, there are few hard and fast rules for assessing the usefulness of such a model. This is especially true of models that result from the application of a variable selection technique. The R and F statistics in these situations may not give a meaningful indication of the model's applicability. More than just a model's statistics are needed if an analyst is to be satisfied that a model will accurately predict the system's cost. By obtaining a good knowledge of the kind of equipment with which he is dealing -- its characteristics, the state of its technology and the available data, the analyst will be able to develop a particular model structure based on sound technological reasoning.

In situations where a variable selection technique is applied to the data to obtain a "best" prediction equation, the permutation test can aid the researcher in the evaluation of this model. It allows the analyst to perform a valid test of hypothesis of the statistical significance of the particular model structure. In situations such as that demonstrated for Subsystem B, where the test indicates that the model is not statistically significant, the data available is such that the possibility of chance correlation is likely. Possible recourses that may be useful:

1. Recheck the definitions used for the parametric and cost data.
2. Collect more observations to improve the data base.

3. Validate any questionable data points that lie outside the expected range of values.

In any event, the permutation test is another tool that the researcher can use in evaluating the suitability of the cost estimating equation.

# CHAPTER VI

## CONCLUSIONS

In empirical model building, unlike confirmatory statistical inference, the situation of working with a given model which possesses certain appealing properties is not assured. The statistical properties of the process under investigation are generally complex and not well understood. The researcher is forced to draw ad hoc inferences from what is often nonexperimental data. In these situations, much of the traditional theory is not valid. In this disseration is has been illustrated that pedestrian use of such techniques as variable selection and transformations may result in models whose $R^2$ values are misleadingly large.

Leamer (12) considers the purpose of the data-dependent process of selecting a statistical model to be "data-mining": using empirical analysis to bring to the surface the nuggets of truth that may be buried in a data set. The researcher has available a plethora of possible statistical computer packages to help bring these nuggets to the surface. To help distinguish precious stones from fool's gold, the researcher must depend on his judgment and experience and the extant statistical theory. In situations where a variable selection technique has been employed, there is a paucity of viable statistical methods to aid in the

assessment of the resulting model. Important methods such as residual analysis and cross-validation have not been considered in this study. Such techniques can offer the researcher valuable information about the model specification. However, these specification checks become less effective when the number of sample observations is small.

The concentration here has been on the investigation of the statistical properties of $R^2$ in such situations. These efforts have resulted in some theory, some informative simulations, and some interesting applications. A statistical test for hypotheses of interest for models resulting from selection techniques has been developed. This result, the permutation test presented in Chapter III, fills the void of valid statistical tests created as a result of the use of data analytic procedures. In situations where the classical F test cannot be used, this test gives the researcher a method for testing the significance of his model. This permutation test is actually an application of an old technique (Fisher's randomization test) in an area of practical importance where theoretical results have been elusive.

# BIBLIOGRAPHY


1. Anderson, T. W. An Introduction to Multivariate Statistical Analysis, New York: John Wiley & Sons, Inc. 1958.

2. Alam, K. and K. T. Wallenius. "Measures of Spurious Multiple Correlation in Best Subset Selection." Unpublished Technical Report. Department of Mathematical Sciences, Clemson University. 1975.

3. Alam, K. and K. T. Wallenius. "Distribution of a Sum of Order Statistics from the Gamma Distribution." Unpublished Technical Report. Department of Mathematical Sciences, Clemson University. 1978.

4. Box, George E. P. "Use and Abuse of Regression." Technometrics 8: 625-629. 1966.

5. Cramer, Elliot M. "Significance Tests and Tests of Models in Multiple Regression." The American Statistician 26: 26-30. 1972.

6. Cramer, Harald. Mathematical Methods of Statistics, Princeton, New Jersey: Princeton University Press. 1946.

7. Daniel, Cuthbert and Fred S. Wood. Fitting Equations to Data, New York: Wiley-Interscience. 1971.

8. Diehr, George and Donald R. Hoflin. "Approximating the Distribution of the Sample R2 in Best Subset Regressions." Technometrics 16: 317-320. 1974.

9. Draper, N. R. and H. Smith. Applied Regression Analysis, New York: John Wiley & Sons, Inc. 1966.

10. Fisher, R. A. "The General Sampling Distribution of the Multiple Correlation Coefficient." Proceedings of the Royal Society of London A 121: 654-673. 1928.

11. Hocking, R. R. "The Analysis and Selection of Variables in Linear Regression." Biometrics 32: 1-49. 1976.

12. Leamer, E. E. Specification Searches, New York: John Wiley & Sons, Inc. 1978.

13. Lindley, D. V. "The Choice of Variables in Multiple Regression." Journal of Royal Statistical Society 30: 31-53. 1968.

14. Miller, Bruce M. and Michael G. Sovereign. "Parametric Cost Estimating with Applications to Sonar Technology." Unpublished Paper. Naval Postgraduate School, Monterey, California. 1973.

15. Mosteller, Frederick and John W. Tukey. Data Analysis and Regression, Reading, Massachusetts. Addison-Wesley Publishing Company, Inc. 1977.

16. Rencher, A. C. and Pun Fu-Ceayong. "Inflation of R2 in Best Subset Regression." Tentatively accepted for publication in Technometrics.

17. Thompson, Mary L. "Selection of Variables in Multiple Regression: Part I. A Review and Evaluation." International Statistical Review 46: 1-19. 1978.

18. Thompson, Mary L. "Selection of Variables in Multiple Regression: Part II. Chosen Procedures, Computations and Examples." International Statistical Review 46: 129-146. 1978.

19. Tukey, John W. "On the Comparative Anatomy of Transformations." Annals of Mathematical Statistics 28: 602-632. 1957.

20. Wallenius, K. T. "Regression Analysis in Parametric Costing and Pricing: Pitfalls, Problems, and Potentials." Unpublished Technical Report. Department of Mathematical Sciences, Clemson University. 1975.

21. Walls, Robert C. and David L. Weeks. "A Note on the Variance of a Predicted Response in Regression." The American Statistician 23: 24-26. 1969.

22. Wilks, S. S. "Determination of Sample Sizes for Setting Tolerance Limits." Annals of Mathematical Statistics 12: 91-96. 1941.

23. Zirphile, J. Letter to the Editor. Technometrics 17: 145. 1975.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|
| 1. REPORT NUMBER: N120 / 2. GOVT ACCESSION NO.: AD-A097536 | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle): On the Degree of Inflation of Measures of Fit Induced by Empirical Model Building | 5. TYPE OF REPORT & PERIOD COVERED: Technical |
| | 6. PERFORMING ORG. REPORT NUMBER: TR #350 |
| 7. AUTHOR(s): T.B. Edwards and K.T. Wallenius | 8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0451 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS: Clemson University, Dept. of Mathematical Sciences, Clemson, South Carolina 29631 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 047-202 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Code 434, Arlington, Va. 22217 | 12. REPORT DATE: 8-8-80 |
| | 13. NUMBER OF PAGES: 68 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | 15. SECURITY CLASS. (of this report): Unclassified |
| | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Regression Analysis, Empirical Model Building, Significance Tests, Parametric Cost Estimation.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This dissertation explores the distributional properties of commonly used statistics developed in the course of empirical model building. A review of some of the more noteworthy efforts to investigate the distribution of the coefficient of determination, $R^2$, in best subset regression is given. To overcome the shortcomings of these results, a permutation test based on Fisher's randomization test is developed to provide a practical basis for assessing the statistical significance of a regression in such situations.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED (OVER)
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. (continued).

An investigation is made into the distributional properties of the multiple correlation coefficient in the choice of a transformation of the dependent variable,y. The study investigates the possibility that pedestrian use of transformations, such as $y^* = (y+c)^p$, may lead to an inflationary effect on the sample correlation.

A practical management science application of the statistical procedures developed in this study is explored in the area of parametric cost estimation.